ELEMENTS OF PROBABILITY

ELEMENTS OF *PROBABILITY*

BY

H. LEVY
M.A., D.Sc., F.R.S.E.
PROFESSOR OF MATHEMATICS,
IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY

AND

L. ROTH
M.A.
ASSISTANT LECTURER IN MATHEMATICS,
IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY

OXFORD
AT THE CLARENDON PRESS

Oxford University Press, Amen House, London E.C. 4
GLASGOW NEW YORK TORONTO MELBOURNE WELLINGTON
BOMBAY CALCUTTA MADRAS CAPE TOWN
Geoffrey Cumberlege, Publisher to the University

First published 1936
Reprinted lithographically in Great Britain
at the UNIVERSITY PRESS, OXFORD, 1947, 1951, from
corrected sheets of the first edition

PREFACE

During the past quarter of a century the subject of Probability has acquired a new importance in science, partly because of the more recent stress on statistical laws in mechanics and partly because of the rapidly expanding use of statistical methods in medical, biological, engineering, industrial, and social problems.

Writers have approached Probability from very diverse angles but little attempt has been made at any sort of unification. At times it is regarded as a branch of symbolic logic, sometimes as a series of empirical conclusions based on experimental practice. Certain writers see it as a branch of pure mathematics, others as a description of a state of mind. To some it is of philosophical, to others of scientific importance.

The authors have taken the view that Probability is an essential of scientific method, and that a probability estimate, however it is approached, has to be seen and interpreted as a guide in scientific procedure. Thus these various treatments are in reality partial aspects of the same topic, where in each case the form of analysis has been decided by the particular scientific purpose for which the treatment has been attempted.

The present book, claiming to be no more than an elementary treatment, makes no effort to cover all these fields. The earlier mathematical portions are restricted mainly to simple considerations of Mathematical Probability and its linkage with Statistics in a form suitable for non-mathematical students; hence the inclusion of the material of Chapters III and IV. At the same time the authors have striven to provide a detailed criticism of the various self-contained theories of probability that have been advanced from time to time. This has compelled them to embark on certain considerations of scientific method and, later in the book, on more advanced mathematical problems in Probability, without, however, entering into fields such as Statistics proper or other branches of physical science farther than has been essential for this purpose.

While most of the examples are new, a number have been selected from Whitworth's *Choice and Chance* and these the authors here gladly acknowledge.

H. L.
L. R.

CONTENTS

CHAPTER I

HISTORICAL INTRODUCTION (p. 1)

CHAPTER II

THE SCOPE OF PROBABILITY

CHAPTER III

THE THEORY OF ARRANGEMENTS

CHAPTER IV

ELEMENTARY THEOREMS ON MATHEMATICAL PROBABILITY

CHAPTER V

BERNOULLI'S THEOREM

CHAPTER VI

EXTENSION TO CONTINUOUS DISTRIBUTIONS

CHAPTER VII

THE THEORY OF ARRANGEMENTS (2)

CHAPTER VIII

THE EMPIRICAL THEORY OF DISTRIBUTIONS

CHAPTER IX

THE USE OF PROBABILITY IN SCIENTIFIC INDUCTION

CHAPTER I

HISTORICAL INTRODUCTION

THE theory of probability arises from a number of different sources. It already manifests itself in certain practices resembling insurance which were known to antiquity; thus, the Roman *collegium* or guild paid a sum of money to the surviving relatives upon the death of a member, a custom which was continued by the medieval guilds. In 324 B.C. a Greek named Antimenes devised the first system of insurance mentioned in history; he guaranteed owners against the loss of their slaves for a premium of 8 per cent. per annum. The marine insurance trade likewise originated in Greek times with the practice of bottomry or sea-loans; when a merchant sent a cargo abroad he received an agreed sum from a banker which he repaid with interest if the cargo arrived safely, but retained if it failed to do so. It seems clear that in such bargains the prevailing rate of interest was high.

The early history of insurance does not appear yet to have been thoroughly explored; that of banking and exchange is, on the other hand, well documented. In the fifth century B.C. banks had already been established in Athens. We know that by the end of the thirteenth century the Italian and, more especially, the Florentine merchants dominated the entire trade of Europe, and that in 1350 they had banking establishments in most of the European capitals; their power was such that they were able to finance wars, control international exchanges, and dictate monetary policy at large. It may be added that at this time a regular rate of exchange began to be quoted in London between English and Flemish currency.

Henceforward financial operations in Europe took on something of their present-day character, including the deliberate policies of inflation and deflation with which we are only too familiar. In this connexion we may note the steps taken by Sir Thomas Gresham, in 1552–3, to restore fallen English credit by pegging the exchange, selling foreign currency in Antwerp, and placing restrictions upon the trade with Flanders. All these operations involved actuarial problems in probability, however rudimentary. The methods of insurance, which date, as we have seen, from very early Greek times, developed without any aid from the actuarial principles with which they are nowadays associated: these latter grew out of a different order of ideas, which we have now to consider.

It is not until the Renaissance that the subject begins to re-emerge in a new setting. During the sixteenth and seventeenth centuries a great deal of the leisure of the European aristocracy was occupied with games of chance and gambling in general. This class did not number among its members any mathematicians capable of handling the problems that naturally suggested themselves, but nevertheless it happened that from time to time problems of chance were passed on to the mathematicians of the period. Perhaps the only exception to this rule was Cardan, himself an inveterate gambler (notorious for his theft of Tartaglia's solution of the cubic) who, somewhere about 1550, wrote a small gambler's manual; the book was not, however, published until 1663. Galileo (1564–1642) had his attention directed by an Italian nobleman to a problem in dice, the solution of which is the first recorded result in the history of mathematical probability.

The problem is as follows: Whereas when three dice are thrown the numbers 9 and 10 can each be obtained in 6 ways (different from each other), yet it is found from actual experience that 10 appears more often than 9. How can this be accounted for? In his work (which did not appear until 1718) Galileo makes an analysis of all possible cases and shows that, of 216 possible ways of throwing three dice, 27 are favourable to the 10 and 25 to the 9. Nowadays we should solve such a problem by the method of Chapter VII; it represents the first successful attempt to explain the frequency of appearance of certain groups of numbers by an analysis of the possibilities that might arise.

Twelve years after Galileo's death a correspondence began between Pascal and Fermat which gave the first real impetus to the theory. The Chevalier de Méré, a French gentleman with mathematical interests, propounded certain questions to Pascal, who communicated them to Fermat. Of these the most important is the famous 'Problem of Points' which in varying forms was to occupy a central place in the theory for the next century and a half. It was first enunciated by Pascal in 1654, as follows: Two players, with equal chances of winning a point, are playing a game for three points. If they wish to break off the game before the end, how shall the stakes be divided? Pascal solves this problem and later enunciates without proof the results for a game for $n+1$ points in the case where one player has already n points and the other none, and where one player has one point and the other none.

Fermat's solution of the problem, given at the same time, is for the case where one player requires 2 points and the other 3 points, to win; his method is essentially the same as that given later in Chapter V. Pascal applies this method to a similar problem in which there are three players. In the same year was printed his *Traité du triangle arithmétique*, which is the earliest treatise on the theory of combinations, and contains, among other things, the familiar formula for the binomial coefficient ${}^nC_r$. Pascal uses the results of this work to solve the problem of points in the case where one player requires m points and the other n points to win.

In all this we see that the setting of the problems is a gambler's one, although both Pascal and Fermat are interested primarily in the mathematical analysis. In this connexion we may note a distinction between the progress of the theory in Catholic and Protestant countries; in the latter the interest was concentrated on quite different topics—thus, Newton, who was born the year Galileo died, seems hardly to have concerned himself with questions of this nature. Almost the sole exception was Huygens who in 1657 produced the first treatise on gaming and dicing problems. This remained the best account of probability until the advent of James Bernoulli, Montmort, and De Moivre, all citizens—at any rate by birth—of countries in which gambling was not frowned upon, that is, in which the Catholic feudal aristocracy was not yet restricted by the rising Puritan class of burghers.

To us the interesting feature of the development of probability at this time is the fact that it began to be cultivated, apparently on a different basis, in England and Holland. These countries were Puritan because the burgher class, the townsmen, had already succeeded in asserting themselves; they were more interested in problems of trade exchange and questions related to the growth of town population. Thus in 1662 we find

Captain John Graunt devising a method for utilizing the weekly returns of deaths in the City of London to determine the growth of the capital, while in 1671 John De Witt published researches on the mathematics of annuities, in Holland. Halley the astronomer published a memoir in the *Philosophical Transactions* for 1693 based on the tables of births and deaths for the city of Breslau during the period 1687–91. He gives a table showing the numbers of the population aged n years, and shows how to find the value of an annuity on the life of a person of given age. He constructs a table of annuities for every fifth year of age up to 70 years; and he considers also the question of annuities on joint lives.

From the end of the seventeenth century to the middle of the eighteenth century was one of the most fertile periods in the history of the purely mathematical theory. During this period James Bernoulli (1654–1705), Montmort (1678–1719), and De Moivre (1667–1754) between them developed the greater part of the elementary theory as it is known to-day, illustrating their work throughout by problems in games of chance, from which it originated.

To James Bernoulli is due an extension of the problem of points; he obtains, substantially by present-day methods, the probability of throwing a given number with n dice; and he solves the problem of the 'duration of play', that is, of finding the probability that a player should win all his opponent's money, given the players' initial capital and their respective chances of winning a point. But his remarkable contribution to the theory is the theorem known by his name (pp. 58–60), the second part of which consists of an approximation to a probability by purely algebraic methods.

The work of Montmort goes over the familiar ground of dice and card problems; in addition it comprises valuable additions to the theory of permutations and derangements, including the solution of the 'problem of treize' (p. 97), and contains the elements of finite differences and the theory of recurrence relations. Many of these results were arrived at independently by De Moivre to whom, moreover, are due the formulae for the chance of throwing a given number with an n-faced die, and that of an event succeeding *consecutively* a given number of times. To De Moivre is due the idea of approximating to probability formulae by means of logarithms; in this connexion he discusses the approximation to the value of the binomial coefficients occurring in Bernoulli's Theorem, and gives a formula which is practically equivalent to Stirling's Theorem (p. 67); it would appear that this theorem had been discovered at about the same time by Stirling himself.

We thus see that, in the effort to discover new mathematical methods to handle problems in probability, there emerged a great deal of work on permutations and combinations, finite differences, recurring series, the idea of summation of infinite series, and many new trigonometrical formulae. These were still in the main a continuation of developments in the Latin countries; the problems dealt with were those that arose from the way of living of the aristocracy. But a new period was setting in, one of criticism and examination preparatory to the French Revolution of half a century later. We can observe the beginnings of this phase in the controversies that arose between Leibniz and James Bernoulli; the latter had attempted, by inverting his theorem on the probability of occurrence of a group of events, to determine the probability of the event itself. Thus what later became a major issue, the 'probability of causes', was raised in mathematical and philosophical form for the first time.

Meanwhile, under the influence of the work of English experimentalists, mathematical physicists, and astronomers, the same problem arose in a new form, one associated with what is called the 'theory of errors', the reasons that can be adduced to explain why sets of observations of the same measured quantity are always, to some extent, discordant among themselves. This was a problem of theoretical science, arising from the needs of experimental practice, and it was one that was certain to intrigue natural philosophers studying scientific laws from a mechanistic standpoint.

From the scientific point of view Thomas Simpson, in his *Miscellaneous Tracts* (1757), was the first to examine critically the implications of taking the mean of a set of astronomical observations of the same event. Thus this theory, now an integral part of the subject of the significance of errors, owes its origin to astronomical needs. Naturally, the French experimentalists were by now equally concerned with the same problem. In 1770 Lagrange published his memoir on the method of taking the best value from among a series of observations.

This work, which had in part been anticipated by Simpson, discusses the probability that the error of the mean of n observations should lie within assigned limits, and determines the most probable error of the mean. Again, if it is known that the errors in a set of observations must be one of the numbers ± 1, $\pm 2, \ldots$, $\pm m$, and that the chances of these errors are equal, or proportional to given quantities, Lagrange shows how to determine the probability that the error of the mean should have an assigned value or lie within given limits.

All these results are obtained by expansion of multinomial expressions and other purely algebraic processes; but at the same time a new conception was introduced by Simpson and Lagrange which proved later to be exceedingly fertile in analysis —the idea of an *error curve*. For reasons to be explained in this book, 'errors' or divergences from the 'true' value necessarily consist of a discontinuous set of data; but apart from the calculus of finite differences, which was still a comparatively new and little known subject, the whole field of mathematics concerned itself with 'continuous' phenomena. Thus, in the face of mathematical limitations, the facts regarding the nature of error were altered to suit, and both Simpson and Lagrange introduced the notion of continuous variation in error. The analogy did not proceed very far; but nevertheless, the concept of errors in a continuum x with a probability function $\phi(x)$ had now found its place.

In 1778 Daniel Bernoulli published a memoir on errors of observations, in which he remarks that the common method of treating discordant observations, by assuming that the true observation is the mean, presupposes that they are of equal weight, whereas small errors are surely more probable than large ones. Bernoulli therefore proposes to measure the probability of an error x by the number $\sqrt{(r^2-x^2)}$, where r is a constant; then the best value x to be obtained from a set of observations $x_1, x_2, \ldots, x_n$ will be that which makes the product $\sqrt{\{r^2-(x_1-x)^2\}}\sqrt{\{r^2-(x_2-x)^2\}}\ldots$ a maximum. In effect Bernoulli thus assumes the probability curve to be a circle and applies to it the method of inverse probability (p. 164).

The idea of continuity in connexion with probability shows itself in other researches of Daniel Bernoulli, in which his purpose is to demonstrate the use of the differential calculus. For example, he discusses the probable distribution of liquid in three urns, initially containing different liquids, if for a time t liquid is allowed to flow from the first to the second, from the second to the third, and from the third to the first. We may also note here the work of Buffon who in 1777 applied the notion of probability to geometrical problems; thus, if a coin is thrown on a table ruled in squares or equilateral triangles, it is required to find the probability that it will fall clear of the bounding lines. Buffon's most famous problem (p. 86), requiring the use of integral calculus for its solution, is found in the same work. It is of interest to note that the result has several times been used to calculate experimentally the value of π with, however, suspiciously good results.

The critical work of the French Encyclopédistes, to which we have already alluded, did not proceed far, conducted as it was by individuals who were for the most part non-mathematicians and who failed therefore to distinguish between those considerations which are mathematically and those which are socially important. Even a distinguished mathematician like D'Alembert, who directed his criticism at the fundamental definitions in probability theory, succeeded only in arriving at the most preposterous conclusions. The Marquis de Condorcet dealt with such questions as the probability of election of a candidate by a given number of voters, and the probability of a tribunal arriving at a true verdict in a trial. In view of his faith in the necessary progress of the human race towards happiness and perfection, it is one of the ironies of history that he himself was condemned by the revolutionary tribunal.

It is during this period that the problem of 'inverse probability', first considered by James Bernoulli, again shows itself, in two posthumous memoirs by Bayes which appeared in the *Philosophical Transactions* for 1764–5. Bayes gives, in geometrical form, the theorem that, if an event has happened p times and failed q times, the probability that the chance of success will lie between the values a and b (all values being equally likely) is $\int_a^b x^p(1-x)^q\,dx \Big/ \int_0^1 x^p(1-x)^q\,dx$. Bayes then proceeds to evaluate these integrals by approximation. It would be interesting to discover whether the investigations of Euler and Legendre on the Beta function $\int_0^1 x^p(1-x)^q\,dx$, which began shortly after 1770, were suggested by the work of Bayes. For us, however, its importance lies in the evidence it affords of the convergence of the subject-matter treated in England towards that of France on the threshold of the Revolution: this, of course, is only a slight aspect. Bayes, himself a clergyman living in the middle of the eighteenth century, turned his attention to these questions, directly or indirectly, under the influence of a sceptic like Hume (1711–76) or an idealist like Berkeley (1685–1753). These latter were themselves working on the ideas of Locke (1632–1704) and Hobbes (1588–1679). Hume, we know, made frequent contact with, and was much influenced by French writers; thus it was in this atmosphere that Bayes attempted to state in symbolical form the relation between cause and effect as it shows itself in probability. It is worth recollecting that, diverse as their outlooks may be on other matters, Locke, Berkeley, and Hume are at one in their distrust of mathematical reasoning and tend to rely on probability rather than on certainty.

If any single person has to be accorded the merit of synthesizing the development of the subject at this stage, that person is Laplace (1749–1827) who, living and working throughout the revolutionary period, drew together the theoretical and philosophical conclusions which had emerged from the problems of gaming on the one hand, and from the discussion of experimental errors, on the other. In addition Laplace established the connexion between these and the corresponding questions in mortality and life tables which lie at the basis of insurance statistics. It is here also that the first specific statement of the Error Function is formulated; and although it was later discovered independently by Gauss (1809) we can accept the view that all the essentials of probability theory and most of the deductions from it are contained in Laplace's great synthesis. From this time onwards it was inevitable that developments in any one of the fields—philosophical, logical, mathematical and experimental, industrial, financial, actuarial and statistical—were bound to affect each other and to grow from the same broad principles. One of these principles, established by Laplace, is the method of Least Squares, which he deduces from a set of very general assumptions. He shows, in fact, that if we suppose the mean of a set of observations to be the most probable value, and positive errors to be as likely as negative ones, the error function for the observations is of the form $\frac{c}{\sqrt{\pi}}e^{-c^2x^2}$.

Actually, the method of Least Squares had previously been used in astronomy by Euler and Gauss, but Gauss was the first who endeavoured to justify it by an appeal to probability theory.

The beginning of the nineteenth century marked a change of profound importance, if not in mathematical methods, at least in the subjects to which these methods were applied. The Industrial Revolution had already set in, with its modern problems of factory production and increasing populations; from these emerged a vast array of social problems which in response to a slowly developing public conscience were becoming the subject of closer and more refined statistical investigation. Thus 1801 saw the initiation of the English population census. A short time later the growing Trades Union movement began to maintain a continuous index of unemployment figures among its members. Simultaneously, under the drive of industrial needs, and with the funds allotted in universities and elsewhere to experimental studies, scientific investigation proceeded apace and with it a whole range of new problems emerged.

In a sense science was, however, largely in the engineering phase, and while questions of experimental error were still discussed, the scientific outlook was highly mechanistic, with little regard for any consideration of statistical qualities in Nature. But the Industrial Revolution, which brought about an immense increase in production, was one of the driving forces towards foreign trade; here, then, on the side of insurance a new impetus was given to the development of the subject, in a field where mechanism had no place and average changes were the qualities that required study. We therefore find during this period a development of those methods of a statistical nature which are required in commercial expansion and social investigation.

Nevertheless, experimental work was proceeding on chemical and physical principles; in particular, interest was focused on the characteristics of gases and gas mixtures, and the pressure laws governing them (possibly under the influence of the new uses for illumination to which inflammable gas was being put). As early as 1660 Boyle had discovered his gas law from entirely experimental considerations; the idea that a gas, impingeing on an obstacle, consists of individual particles, and that the pressure it exerts results from this mutual impact, had been noted many centuries before, and various unsuccessful applications of the idea had already been made by Newton. In 1738 Daniel Bernoulli showed that Boyle's Law follows from the hypothesis that the gas consists of a large number of moving particles, and that the pressure arises simply from that exerted by the gas on the walls of the containing vessel. So the position remained until the turn of the century when, as we have indicated, attention was drawn to the properties of gas mixtures. Thus, in 1802, Dalton enunciated his law for the pressure of gas mixtures, basing it on the tacit assumption that the motion of all the particles involved was uniform. By the middle of the century Clausius (1847) and Joule and Kronig (1857) had shown how to express the pressure in terms of the mean velocity of the gas particles.

Meanwhile, the philosophic problems associated with probability, which had emerged from the writings of the Encyclopédistes, were being examined and extended by De Morgan, Venn, Boole, and others. The law of Laplace-Gauss was well accepted as the necessary distribution function for a combination of 'random' factors. By 1860 Maxwell was therefore in a position to apply these ideas to the random motions of gas molecules, and from this there rapidly developed an elaborate statistical theory of gases.

We should note that this marks a culminating point in the theoretical development, in the sense that we have presented a new class of problem in scientific method. For, by his analysis, Maxwell showed how the characteristics of a large mass and the laws exhibited by it in various circumstances are related to the corresponding characteristics of particles at a 'lower' level. Although since that date many fruitful developments of Maxwell's theory have occurred, the next stage in its application was not until the beginning of the twentieth century, when the experimental discovery of still more elementary forms of matter (electrons, protons, neutrons) threw up a similar type of problem for study: namely, how to express the characteristics of the atom or molecule in terms of the more elementary characteristics of the electron, proton, etc., on the assumption that these show themselves as the result of statistical combination. We may put it shortly by saying that the step from Newtonian theory for the motion of a body to Maxwell's theory for the characteristics of a gas is similar in type to the step from atomic characteristics to the quantum theory.

It remains to point out that there exist at the present day groups of investigations of a statistical nature arising from insurance, actuarial analysis; industrial statistics and their application to production and distribution; the statistics of new social problems and the statistical approach to questions in purely scientific inquiry, including genetics, quantum mechanics and mathematical logic, and these, admittedly requiring specific treatment, are usually dealt with as if they were separate and distinct fields. All these developments require a new unification and synthesis, such as was performed by Laplace in his day; the efforts that have been made to this end, merely by the production of a theory of probability as an extended branch of logic instead of as an actual and vital part of scientific process, must, when seen in perspective with this historical movement, fail in their function. That unification has yet to be found.

CHAPTER II

THE SCOPE OF PROBABILITY

1. The meaning of chance

ALL events in the universe are interrelated and affect each other to a greater or less degree; for example, the reader of this book will be affected by all the factors which brought the book into existence, and these range from the manufacture of paper and ink on the one hand, to the history of the authors, their parents and teachers, on the other. Thus all events have an enormous number of causes, some more important than others. It follows that, in any attempt to obtain information about them, some selective principle is necessary in order to eliminate what we suppose will turn out to be the less relevant facts in a particular case; indeed, by the mere use of the word 'event', we are focusing our attention on the thing that interests us, all other things being for the moment irrelevant.

Science is concerned with particular kinds of events which interest us. The procedure which characterizes scientific method consists in isolating *rational* sequences of events, that is, events which appear to form a logical chain when interpreted in the light of certain fundamental assumptions. Thus, a ball is projected into the air with a specified speed: it rises to a certain height and reaches the ground at a certain distance from the point of projection. A scientific study of this projectile attempts to connect this sequence of events so that one or more of them follow as a *logical* conclusion from the others. For this purpose, in the first place we ignore all other events except these, e.g., we ignore the temperature of the atmosphere, the possible defects in the apparatus used for the projection, and the personal views of the experimenter; and in the second we assume the operation of some guiding principle, frequently described as a 'law of force'. Such a problem belongs to the science of *rational mechanics*, which by postulating laws of force purports to deduce mathematically the effect of a given system of forces acting on a given system of bodies. In other words, one of the aims of mechanics, as of any other branch of science, is prediction. (What interests us, in a sequence of events, is the way in which they can be grouped together to facilitate prediction and thus effect control over Nature.) The accuracy of the prediction will consequently depend not only on the selection of events, but also on the guiding principle which we were led to make in formulating our science. We are not justified in the first instance in assuming that this process will lead to results which agree with observed facts; if, however, we wish to sharpen the accuracy of our prediction, it is clear that we can do so by making a study of those events which we rejected previously as being less relevant to the problem. This might also necessitate a change in our guiding principle. This sharpening process may be repeated again and again. At any stage *we define the difference between the event predicted and the actual observed occurrence as a chance effect.* While one field of science, which we have called a rational system, occupies itself with predictions which involuntarily exclude or ignore these chance differences, another field takes them as its object of study, under the name of 'deviation' or 'experimental error'. It is with this that the calculus of probability is concerned in its application to experimental practice.

For certain purposes of analysis, a guiding principle is here again frequently assumed, when 'chance' is conceived as itself the result of a large number of equal elementary causes combined together. In the examination of this theory, points are often illustrated by models and analogies such as those dealing with balls chosen from urns, each such choice being thus regarded as a simple, elementary event.

We must begin our study with a word of warning. The abstract theory of probability, which seeks to comprehend those facts which elude the ordinary rational systems, must itself of necessity be a rational system, working by mathematical methods and based on certain assumptions. So it frequently happens that problems which appear to be about physically real things, such as balls extracted from an urn or a coin tossed in the air, have nothing specifically 'real' about them, in relation to balls and urns: they are simply abstractions fitted into a picture to assist the mathematician. The justification for using such abstractions in our problems cannot rest finally on any theoretical basis alone, but in the last analysis has to be found from the experimenter before or after the abstractions have been applied. In any case we must distinguish between the mathematical *problem* of choosing a mathematical ball from a mathematical urn—an imaginary problem—and the *actual* urn, the balls contained in it, and the actual process of choice. The former may guide us in analysing the latter.

An example will make the need for this distinction clear. If 100 persons each have to choose a number between 0 and 9 inclusive, how often will the numbers 0, 1, 2,... be chosen? The abstraction which a mathematician *might* make from this problem would leave him with a purely mathematical question concerning arrangements, the answer to which is, that each of the numbers will 'probably' be chosen ten times. But this is not the real question; what we want to know is how people *actually choose*, and here we are faced by considerations of a psychological and social nature. In point of fact it has been found by actual testing of a large number of individuals that 7 and 3 are much more frequently chosen than any other number; these numbers both, of course, have a long historical and religious tradition behind them. As we see from such an example, the question whether the abstraction may be validly applied in a given case is not to be begged. The mathematical problem deals with the number of arrangements that can be conceived as possible in the circumstances, the physical problem with the groups of these which actually come into play. We can develop a mathematical theory of arrangements but a separate justification has to be found for it if it is to have practical applications. Thus, the mathematician may postulate that 'an event can happen in two different ways'; whereas the physicist knows that it does happen in one way only.

In the above problem we recognize two questions inherent in the theory of probability: a mathematical question concerning possible arrangements, and a physical question concerning actual choice or action. There is also a third kind of problem which we now consider. Most human beings, even if they are not scientists, analyse events in a rational way, that is, they recognize *order* and *recurrence* and are so led to develop a sense of expectation as a subjective reaction. If we study a person scientifically we may ask whether his expectation of an event is *justifiable*, that is, whether his past experience is sufficient to produce the expectation that would correspond to the reality which the future will bring forth. For instance, if one wakes up in the morning and hears a cart rattling in the street, there comes the thought, 'I expect that is the milkman', or else 'It is probably the milkman'. Obviously it is either the milkman or it is not, and it is one's past experience, in which one's expectation has sometimes been verified and sometimes not, that determines the *strength* of the expectation. Whether that expectation will be verified or not will depend on how far our psychological reactions conform closely to the underlying processes of the external world. We see, then, that such a question is not to be decided by a study of all the possible arrangements which the future may conceivably bring forth: we cannot thus be sure, without elaborate investigation, that psychological expectation is itself a sure guide to future occurrence.

To sum up: in our analysis of situations relevant to 'probability' we have discovered three possible fields of study, all in some way interrelated and each a partial approach to the general problem:

(1) a mathematical theory of arrangements;

(2) the frequency of actual occurrences;

(3) the psychological expectation of a participant.

Problem (2) is the one which arises in actual practice, when in describing the course of past events we attempt to predict the future: in this respect it does not differ from every other experiment, which is always concerned with the past as a guide to the future. Problem (1) is a mathematical discussion of abstractions which may be useful in (2) if they are shown to be relevant; while (3) represents the subjective state of a person who possibly makes a rough use of (1) and (2) when he is faced with the events in (2).

In (1) the conception and practice of *chance* do not occur: every problem must be precisely defined and has a precise answer. For example, we may ask, out of a pack of 52 cards, what proportion of all possible groups of 13 will contain 4 aces? Here no question of chance arises. In such a problem the exact number of cards, and the kind of hand, are specified: there are no ambiguities in the situation—the 52 cards and the 4 aces are *isolated* in an abstract way from all the rest of the universe: in short, they are *given*. Any actual process of selection is deemed irrelevant, and the answer is unique. Precisely the same situation arises with a geometrical problem: thus, we are given a triangle with certain properties and we proceed to deduce certain consequences. On the other hand, chance, as we have defined it, enters into (2), and again in (3), since the individual concerned makes his own analysis which is necessarily partial; but what is chance to him need not be chance to the scientist engaged with problem (2).

Chance in Scientific Observation

A scientific observation depends not only on instruments but on the circumstances in which they are used—for example, the individual who performs the experiment, the temperature of the laboratory, and so on. Hence the results depend, to some extent, on the differences between individuals. The object of all scientific experiment is to obtain objective information about the world: by objective information we mean information that can be stated in a form independent of the particular experimenter and his idiosyncrasies. We call this information *invariant* to the individual.

Suppose that we wish to measure the length of a desk: whatever definition of 'length' we may adopt, if it is to be of any use for scientific purposes it must be invariant to the observer. But one observer applies a measuring rod to the desk and finds that it records 25·1 inches, another finds instead the reading 25·2 inches, a third 24·9 inches, etc. What then is the length of the desk? At the end of such a series of observations a scientist has in his possession a set of *numbers*, which represent all the measurable information that he can obtain for his purpose. He has then to say to which, if any, of his numbers the term 'length' will be applied; the differences between the selected number (the 'length') and the rest he assigns to 'chance'. They are presumably due, among other things, to the observer who, so far as an invariantive statement is concerned, is a chance one, irrelevant to the issue. The chance differences are said to be 'errors of observation'; but in effect such a term is simply a means of grouping together all that remains after the rational abstraction, which has been called 'length', has been made. In this way the idea of chance becomes identified with the cause of so-called 'experimental errors': the one implies the other. The definition of length really specifies the method of isolating the experiment from the rest of the universe in an attempt to obtain objective information and to build up a logic of science: the 'errors' represent the real connexion (or part of it) between the isolate and its residue with respect to the universe.

From the illustrations we have given it will be observed that the difference between the mathematical and the physical approach to a problem is that, whereas in the former the field of discourse is defined in advance, in the latter the primary object of our inquiry is to find it. A physicist who is studying the properties of matter discovers that it can be broken down into electrified particles; thus he has now found a field of investigation. The mathematician can now begin his analysis with the statement: *Given* two *isolated* electrified particles interacting in a *given* way, can their future behaviour be predicted? Such behaviour can then suggest a new field of investigation to the experimenter who, unlike the mathematician, is never 'given' two isolated electrified particles.

Thus in the one case a mathematical field is postulated and we examine its logical implications: in the physical problem the make-up of the world itself is the unknown, and the object is to discover what in fact is its structure. In practice, however, both physicists and mathematicians work hand-in-hand and supplement each other, as shown in the above example. The subject of probability, therefore, to be complete, has to play its part in both fields; the mathematician has to forge an instrument which the experimenter can use in practice.

2. On the definition of probability

Definition of Mathematical Probability

We propose in the first instance to define 'probability' in a purely mathematical sense, that is, in connexion with problem (1). The definition we give is the following:[†]

'If there is a group of N letters consisting of n_1 letters a_1, n_2 letters a_2, ..., and n_r letters a_r, the probability of a letter specified

[†] See also Peano, *Rend. Accad. Lincei* (5), 21 (1912)$_1$, 429.

as belonging to the whole class $a_1, a_2, \ldots, a_r$ being a letter a_s is n_s/N.'

Having posed this definition we may legitimately ask whether it may be applied in a particular problem, that is, whether the definition of probability has any relevance to an actual experimental case. For instance, if a penny is tossed, what is the probability of a head? We can construct a model problem which we imagine to be like it by making correspond the word 'head' to the letter a_1 and the word 'tail' to the letter a_2, whence we obtain a mathematical solution; we replace the real penny and the action of tossing by two arrangements which we may call either 'head and tail' or 'a_1 and a_2': in this way the actual penny no longer concerns the mathematician.

But this gives us no definition of probability of a physical event, such as the tossing of a coin: our knowledge of such an event implies knowledge of the circumstances in which the coin is tossed. An experimenter who studies the problem might ask himself how frequently the head appears: he might study the detailed process of tossing but whether he is entitled to make a precise prediction on that is another matter. An onlooker might, if interrogated, reply that heads and tails are *equally likely*: his answer emerges from a collective experience, a result of having seen actual pennies spun. To bring the term 'equally likely' into a mathematical definition would be to confuse (3) with (1), just as to use the term 'equally frequent' would be to confuse (2) with (1).

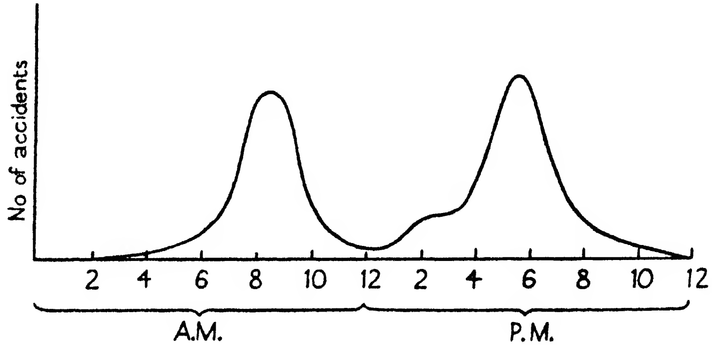
Definition of Statistical Probability

We define a class of event by a distinguishing quality of that class, e.g. the event known as traffic accidents; these grow in number with time and will be referred to as a *population* of traffic accidents. At any given moment the ratio of fatal cases to the total number has a certain value which itself in general varies with time; this ratio we call the statistical probability of fatal accidents. Its importance lies in the practical fact that it is used either as a guide to prediction concerning the number of such cases in the future, or as a factor in determining how we can attempt to diminish them.

We note two points of difference between this definition and the preceding. In the latter the population of events whose arrangements we considered was in general finite and determined, and the probability of any subclass was a matter for deduction. In the case of statistical probability, both the population and the subclass, although defined in Nature, are not initially bounded in extent and, in fact, they grow with time. The significance of the probability lies in its application to future members of this growing class; thus the application is essentially of an inductive nature. This is not to say that the two methods of approach have nothing in common; when we come to discuss what is called the significance of a statistical probability it will be found that the mathematical definition affords us an idealized standard against which the significance may be measured. Statistical probability finds its application in many branches of insurance, in the analysis of demographical statistics, and plays a part in such natural phenomena as meteorology, where the deductive methods of physical science are not yet sufficiently precise to enable satisfactory predictions to be otherwise made.

A priori Probability

There is a form of statistical probability which appears in the literature of the subject under the name of *a priori* probability. Let us suppose, for example, that we are examining the probability of an individual being killed by traffic in the streets of a busy town. Although the actual data from which the statistical probability curve could be drawn are not available, it is nevertheless possible from general considerations based on our knowledge of the circumstances and the impressions we have gained from others' experience, to construct a probability curve which will at any rate serve as a first approximation to the truth. Thus we know that between 8 a.m. and 10 a.m. many people are in the street on their way to work, and that between 4 p.m. and 6 p.m. they are returning home. Moreover, we may expect that the ordinary traffic of the day is also augmented during those periods by the cars belonging to business men. Accordingly, most people would agree in producing a curve like that on the following page. From this we can determine an *a priori* probability; its significance lies in the fact that, if we wish to use it, it gives us a first criterion for judging whether a batch of fatal accidents occurring, say, between 2 and 3 in the afternoon can be regarded as normal or not. Thus it enables us to make a first rough estimate of the probability of obtaining such a sample.



It is, of course, admitted that the details of the probability curve in the figure will vary with the person who constructs it, but there are cases in which no such differences arise. For example, if a penny is tossed, all will agree that on the basis of past experience the *a priori* probability of obtaining a head is $\frac{1}{2}$. This does not rest solely on the mathematical ground that a penny has a head and a tail, but on the additional fact that pennies do indeed, on the average, fall with equal frequency on head and on tail. If the general experience of tossing coins were sufficiently exact and had shown that in fact heads appeared 51 times in a 100, the *a priori* probability would be accepted as $\frac{51}{100}$. We shall see later, when dealing with Bernoulli's Theorem on the mathematical probability of obtaining certain proportions in a given sample, that a knowledge of the proportions in the original population is essential for the solution. There we shall refer to it as the probability of an individual *member* of that population; but in applying the conclusions to samples drawn from it we must bear in mind that the probability in question is merely a precise form of the *a priori* probability which we have been considering here.

Probability as a Branch of Logic

The subject of probability is approached by many writers from a different angle, viz. as an extension of a branch of logic. A set of facts, called the 'data', are stated, and a proposition referring to them is set alongside them; among the numerous relations that might be stated between the proposition and the data we consider one type in particular. While we usually assert that the proposition is either a true or a false statement about the data (it is certainly true if the data *imply* the proposition), an intermediate state may be considered. A class of 30 children may be the data and the proposition, 'All these children have brown eyes'. In this illustration the restricted form of the data tells us nothing about the children's eyes: the proposition is therefore not implied by the data, but is nevertheless not inconsistent with them. If further information were available it is possible that the proposition might be true; but as it stands, it outstrips the data. When such a situation arises it is said that the proposition has a 'probability relation' with respect to the data; the probability relation is then regarded as a member of a class of relations, the extremes of which are 'true' and 'false'. We may say that

'A proposition is true', or
'A proposition has a probability', or
'A proposition is false'.

It will be noticed that this approach to probability suggests that it is primarily psychological; if it were purely logical there would be no escape from the position that the proposition is either implied or not implied by the data. It is when the proposition and the data are not thus rigorously bound together that the psychological attitude enters into the question. We feel that although the implication is not logically complete, nevertheless if *further data were available* the proposition would be found to be true. Thus the probability relation implies that when the proposition is used for enlarging the data it may be found to be true; this views the probability relationship as a step towards the accumulation of further data and the final establishment of a truth or falsehood: otherwise it remains artificially separated from its function.

Consider the above illustration: to say that there is a probability that the 30 children all have brown eyes is futile unless we go on to discover whether they have, or what proportion of them have brown eyes. When this step has been taken, the final data imply the truth or falsehood of the original proposition.

This interpretation of a probability relation gives it a value in scientific method. If, however, we attempt to give it a value in itself by isolating it from the necessary part that it should play in scientific method, the subject may be indeed developed further but necessarily not on the present lines. To appreciate this fact we must return to the concept of expectation: given a set of data and a proposition which outstrips them, each individual, on the basis of his past experience, has a sense of expectation that, if further data were accumulated, the proposition would be verified. A group of experimenters of wide experience in the particular field, i.e. on the basis of previous data not here specified, would presumably agree that they *strongly suspected* or *rather expected* the proposition to be true, or *thought it might* be true. They may thus find themselves agreeing that a gradation in the sense of expectation is associated in their minds with the possible truth of a proposition. To proceed further along scientific lines some objective measure of expectation must be found, otherwise the theory as so constituted cannot come within the range of physical science. It is possible that the expert psychologist might find such a measure, by examining the reactions of the experimenters, but not directly from the data. A statistician might find such a measure, but he would derive it from the data alone, and not from the experimenters' sense of expectation.

An attempt to overcome the difficulty respecting the non-metrical nature of probability when approached in this way has been made by laying down the following axioms:†

'1. If we have two sets of data p and p', and two propositions q and q', and we consider the probabilities of q given p, and of q' given p', then ... the probability of q given p is either greater than, equal to, or less than that of q' given p'.

2. All propositions impossible on the data have the same probability, which is not greater than any other probability; and all propositions certain on the data have the same probability, which is not less than any other probability.'

† Jeffreys, *Scientific Inference*, ch. ii. A very similar artifice is adopted by F. P. Ramsey (*Foundations of Mathematics*, p. 158) but he retains a subjective criterion for the strength of a belief, so that his symbols have an entirely personal reference. See footnote, p. 27.

Here by the phraseology used the sense of expectation has been given a status that is invariant to the individual and is attached to the objective situation; at the same time it is implied that this psychological probability is measurable. To circumvent such difficulties, what amounts to a verbal artifice has been adopted. The gradations in psychological expectation are identified with the real numbers, using instead of the word 'truth' the word ONE, and writing it as 1; and using instead of the word 'falsehood' the word ZERO, and writing it as 0. By the use of this verbal method with the foregoing axioms all probabilities of this nature apparently become measurable by numbers lying between 0 and 1: thereafter it is a simple matter to derive the ordinary formulae for mathematical probability by setting out a series of theorems, such as:

'If several propositions are mutually contradictory on the data, the number attached to the probability that some one of them is true shall be the sum of those attached to the probabilities that each separately is true.'

In this treatment the idea of psychological probability has been transformed merely by use of an analogous terminology into mathematical probability; the fact that psychological probabilities have been stated as numbers, which are additive and range between 0 and 1, would, if these statements were true, imply an elaborately detailed knowledge of psychological processes and their measurable qualities. In point of fact, of course, no such data are available. It follows that, after these assumptions have been made, the subsequent treatment of the subject cannot differ in essentials from that of ordinary mathematical probability; although the fact that it is artificially based on psychological ideas may have the effect of confusing the later interpretations. If it is necessary at all to emphasize the gravity of the assumption that psychological probability is measured by numbers lying between 0 and 1, it is, for example, sufficient to point out that one could equally well arrange that 'truth' should correspond to the colour blue, and 'falsehood' to red, all intermediate colours in the spectrum being assumed to correspond, somehow or other, to intermediate states of feeling. Such an arrangement would imply the same type of fallacy even though, as it stands, it does not immediately involve assumptions of measurability, but merely those of correspondence.

At this point it is desirable to add that, in refuting a purely psychological approach to probability, we are far from denying that that line of development is necessary. We have already said that the concept of probability itself marks a useful stage in scientific method—'useful' in the sense that it suggests the direction in which to seek and interpret data; it is the stage intermediate between partial ignorance and experimentally sufficient certainty.

The Principle of Insufficient Reason

In this connexion it is worth considering a method which various writers have evolved in order to arrive at an estimate of the *a priori* probability. It is commonly stated that if there is insufficient evidence to justify a probability assertion, the latter can be established by referring it to the 'principle of insufficient reason'. Let us quote Jeffreys† on the subject:

'How do we assess the probability of a proposition before we have any means of knowing whether it is true or false? It has often been said that assessing a probability implies some knowledge, and that therefore we cannot assign a probability when we are in complete ignorance. This opinion must be directly contradicted. Complete ignorance *is* a state of knowledge . . . and the probabilities assigned upon it are perfectly definite. If we have no means of choosing between alternatives, the probabilities attached to those alternatives are equal.'

To adopt this standpoint is to deny the whole basis of science. Science is based on knowledge, if only partial, and nothing whatsoever can be built on ignorance: without data no conclusion can be drawn. If the fundamental question of our subject can be stated in the form, 'Given certain data in a given situation, what precise deduction can be drawn from them?' then the problem of drawing a deduction from no data does not fall within its scope. If we are in complete ignorance about an event, then we are in complete ignorance of how to estimate its probability. In this case the principle of insufficient reason asserts that the probability of its happening is ½, since the sum

† *Scientific Inference*, ch. ii.

total of our relevant knowledge may be stated in two mutually exclusive propositions which exhaust all the possibilities:

The event happens.

The event does not happen.

But we must not, however, confuse the nature of the event and the data concerning it with these two verbal propositions. By hypothesis, we know nothing about the event; the above two propositions provide data for another problem, which may be stated simply as:

'Given that a certain statement belongs to a class of two statements, what is the probability that it is the first of these?'

If the principle of insufficient reason is used in this way, it tells us something about the arrangement of statements but cannot provide us with any estimate of truth-probability of their *content*.

It might be maintained that in practice the principle is used in this way to assess the probability as $\frac{1}{2}$ and to base action on the assessment. As an unqualified statement this is definitely untrue; when we are unable to estimate a probability, we may as a matter of convenience *assume* a tentative value of $\frac{1}{2}$, but only as a matter of convenience in practice. Every illustration which can be produced, however, in which the principle appears to provide us with an estimate of probability in the sense stated above, turns out to be so constructed that by *definition* all relevant information that any one would know or immediately seek to discover is automatically excluded. Action is never taken on the basis of no information, and judgement, when it has to be applied, must be applied to some content of fact.

As an illustration of an *abstracted* problem consider the following:

AB is a line of unknown extent, XY is a segment of AB, of unknown extent and position. If P is a point situated in AB, what is the probability, we ask, that P lies within the segment XY? On the basis of the above principle the answer would be $\frac{1}{2}$. There is in reality no such answer, for we have insufficient data on which to make even an estimate of the probability, since the points A, B, X, Y are known only to exist on an infinite line.

We can, however, state the three mutually exclusive propositions which together exhaust all the possibilities:

P lies to the left of X;
P lies in XY;
P lies to the right of Y;

and the probability that a statement which is known to be one of these three will be the second, is $\frac{1}{3}$. In actual fact, no rational being would use such an estimate if, say, he were attempting to recover an article lost in a street AB in which XY was a very small section—even if it were the most brightly illuminated section.

If indeed probability is to be used as a guide to action, as it must be if it is to play its part in scientific method, then the above illustration brings out the weakness of this approach. On this basis, the probability of finding the article in XY would be $\frac{1}{3}$, whether the lamp is present or not; nevertheless, most people would proceed straight to the lamp, since its presence is more relevant to action than any abstract estimate of probability based on mere verbal propositions. It seems clear that when a situation arises in which *a priori* probability can be estimated only by means of the principle of insufficient reason, this probability itself becomes insignificant as a guide to action, and other factors become much more important.

Other Definitions of Probability

In the light of the above discussion, it is worth while examining the definitions that have been given by other writers, as a preamble to their mathematical treatment of the subject.

James Bernoulli begins by defining probability as the measure of the strength of our expectation of a future event: this is clearly a case of (3), and Bernoulli's treatment must, if consistent, lead to a mathematical theory of *psychology*. In spite of his initial definition, his analysis is carried through as if based on the definition (1) and his treatment becomes that of purely mathematical probability.

According to J. M. Keynes,† probability is not concerned with events other than judgements or propositions; thus his treatment, although symbolical in form, is one of a non-measurable

† *Treatise on Probability* (1921).

logic and rules out mathematics altogether in the accepted sense.

J. S. Mill,† quoting from Laplace, says: 'Probability has reference partly to our ignorance, partly to our knowledge. . . . The theory of chances consists in reducing all events of the same kind to a certain number of cases equally possible, that is, such that we are *equally undecided* as to their existence; and determining the number of these cases which are favourable to the event sought. The ratio of that number to the number of all the possible cases is the measure of the probability. . . .'

This is the definition to which Mill himself inclines, and is a confusion of at least two of our three concepts of probability; the confusion is complete when later Mill adds that 'we must remember that the probability of an event is not a quality of the event itself, but a mere name for the degree of ground which we, or some one else, have for expecting it. The probability of an event to one person is a different thing from the probability of the same event to another, or to the same person after he has acquired additional evidence. . . .'‡

From what we have already said it should be clear that Mill's definition does not disentangle the various elements which enter into probability. For he is obviously thinking of (1) when the events are presumed, and of (2) when they are being formed in experimental practice. We have seen how important it is to distinguish between these two concepts; they are not interchangeable although they may be mutually helpful. To take the statistical definition, viz. the actual ratio of favourable, to the total number, of cases from a block of similar past events, as identical with the mathematical definition of probability would be to identify a number, which in general varies with the growing population, with a unique mathematical value which emerges from the definition of certain classes.

The various types of probability estimates may be illustrated by the experiment of tossing a coin. We may say, as has already been suggested, that the *a priori* probability of a head appearing is $\frac{1}{2}$, a number drawn and posited from a wide but unspecified

† *A System of Logic*, 8th edition, Book III.

‡ Cf. Jeffreys, op. cit., p. 10: 'A proposition . . . has one and only one probability. If any person assigns a different probability, he is simply wrong.' See also footnote, p. 22.

past experience. We may say that the mathematical probability is ½ on the grounds that there are only two possibilities, head and tail, and that these are *defined* as having equal probability. Or we may actually perform an experiment; thus in the following table we give the results of tossing such a coin 100 times, and the number of heads recorded after 10, 20, 30,... tosses. It will be seen that the statistical probability ranges from 0·46 to 0·65, and is therefore a function of the size of the population.

No. of heads	*No. of tosses*	*Ratio*
6	10	0·60
13	20	0·65
16	30	0·533
21	40	0·525
23	50	0·46
28	60	0·466
35	70	0·50
43	80	0·537
49	90	0·54
55	100	0·55

It is thus seen, even at this stage, that yet another problem suggests itself as of importance in interpreting such data as those given above. If we associate the mathematical definition of the probability of obtaining a head (namely ½) on any one occasion, with the statistical probability as here defined, we may inquire what is the mathematical probability that in the first 100 tosses of a coin (probability of a head = ½) fluctuations from ½ of this magnitude will occur. We shall deal with this question in Chapter V; but for the moment it is important to recognize how *mathematical* probability may be used to interpret a fluctuating *statistical* probability.

This fluctuation is, of course, necessarily associated, as in the case of a coin, with the method of tossing. It is clear that with a given coin which is tossed by some mechanical process (beginning always with, say, the head upwards), it could be arranged that the result of each toss is always head or always tail; or, alternatively, that the ratio of the number of heads to the number of tails takes a certain series of values within a specified range.

The above example illustrates the fact, which we shall encounter frequently, that in any physical process to which probability is to apply, there are three interlocked elements:

(1) a 'population' P, in the above case, of heads and tails;

(2) a process of selection S (here a mode of tossing);

(3) a sample s drawn from P by the application of S. This process may be stated symbolically in the form $s = S(P)$.

In the previous examples where the coin has been tossed 100 times, it has been shown that, with the particular form of S used in the experiment, the sample s drawn contains numbers of heads in ratios lying between 0·46 and 0·65.

Some discussions of statistical probability, when they attempt to link it up with mathematical probability, try to do so by asserting that the ratio obtained by sampling a population can be made to lie within increasingly narrower limits merely by lengthening the process S. It seems clear, from what we have said, that it is not simply the length but also the *form* of the process that is of importance. The gap in the discussion will not be bridged until it can be shown that there exists some kind of process S which is capable of mathematical and empirical definition, and of leading to such a result; any particular process of this type could then legitimately be called a 'random' one, and the class of such processes would in such circumstances identify the mathematical with the statistical definition. That all processes S do not fall within this category is obvious from the fact that S can be deliberately designed so as to violate the required condition.

The reader is advised to try this experiment himself, and to note that the ratios he obtains are different from those given above.

In his discussion of the subject, Coolidge† attempts to surmount the breach between the mathematical and the empirical approach (i.e. between (1) and (2)) by the following 'empirical assumptions' of the type to which we have referred.

'1. If an event which can happen in two different ways be repeated a great number of times under the same essential conditions, the ratio of the number of times that it happens in one way, to the total number of trials, will approach a definite limit as the latter number increases indefinitely.

† *An Introduction to Mathematical Probability* (1925).

2. If an event can happen in a certain number of ways, all of which are equally likely, and if a certain number of these be called favourable, then the ratio of favourable ways to the total number is equal to the probability that the event will turn out favourably.'

The first of these assumptions is devoid of mathematical precision; first, because the question is begged by the phrase 'the same essential conditions'. This is a phrase which commonly occurs in all branches of mathematical physics. It is often posed as a fundamental proposition in scientific method in the form, 'The same experiment always produces the same results when carried out under the same conditions'. For our purpose it is important to note that *no* two experiments can be the same; invariably they differ in time or place, and almost invariably in experimenter and apparatus. This criticism applies also to the phrase 'the same conditions': the test for 'sameness' in two cases is provided by the *results*, for these are numbers which can be checked against each other. In the last analysis the test whether these conditions have in fact been fulfilled lies in the concurrence of certain intermediate and all the final results. Thus the proposition quoted is meaningless; it represents an effort to abolish a vital distinction between two concepts which differ fundamentally and is simply a concession to mathematical convenience.

So much for the criterion of sameness in the first empirical assumption; in the second place, the assertion that the ratio approaches a 'definite limit' cannot be justified by any mathematical definition of a limit. It has to be dealt with in the manner already indicated.

The second assumption is not an assumption at all, but a definition, as is indicated by the phrase 'equally likely'. This is either an appeal to subjective psychology (under (3)) or a *petitio principii*, in that the measure of the probability, as defined, will by its consistency indicate a criterion for 'equal likelihood'.

An interesting attempt has been made by Mises to erect a theory of probability that would bridge the gap between the classical mathematical and the statistical approach. The former, as we have seen, is concerned with a *given* population and confines its questions to those relating to the relative frequency of various arrangements of the elements of that population. Mises's theory is in the first place a statistical one. He confines his attention to the infinite succession of unit samples as they are drawn from an unknown population or as they are created by the repetition of a particular action (e.g. the tossing of a coin), and the questions that are raised concern the nature of the predictions that can be made regarding the occurrence or non-occurrence of a particular kind of sample in the sequence.

Let the symbols 1 and 0 be used to represent 'success' and 'failure', or 'black' and 'white', the two possible outcomes of an action. Then Mises is concerned with a collection of the type

$$1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1,\ 0,\ 0,\ 0,\ 1,\ 0,\ldots$$

and proposes to define its structure in such a way as to provide a reasonable meaning to the phase 'The probability of the occurrence of 1'.

The nature of the definition of structure, however, must not be such as to destroy the 'random' occurrence of the 1's with respect to the 0's; in other words, there must be present a persistent disorder. This implies, of course, that by no detailed study of the system should it be possible, for example, for a gambler to discover a pattern or law in the occurrence of the symbols of such a form that he could arrange his gambling with any certainty on the occurrence, say, of a 0 or a 1 at a series of allotted positions.

To fulfil these requirements the sequence is restricted by the following two conditions:

(1) If in the first n symbols, there occur m of the type 1, then the sequence is such that

$$\lim_{n\to\infty} \frac{m}{n} = p.$$

The probability of the occurrence of 1 in the sequence is defined as p.

This statement may be put in a form more usual with the treatment of sequences; thus corresponding to *any* small number ϵ it is possible to find a number of terms N, beginning from the left, and a number p, such that for all values of $n \geqslant N$ the ratio m/n will continue to differ from p by less than ϵ.

(2) The second condition is the Principle of Disorder. It demands further that the sequence shall be of such a nature that by whatever system or law, related to the order of the terms, a new sequence be formed from all or some of the terms of the original, all such derived sequences shall separately satisfy the previous condition with the same value of the probability p.

At this stage two points should be noted. Condition (1) at first sight bears a very close similarity to assumption (1) of Coolidge (p. 29).

Here we may remark, however, that whereas the latter took the statement of convergence as an empirical assumption applicable to real statistical data, in the present case the condition of convergence is merely a restrictive property of the collection to be considered. The question whether sequences satisfying such a condition do embrace actual statistical data empirically derived remains open.

The second criticism may in a sense be much more serious. The Principle of Disorder, applicable as it must be to every systematically derived sequence, must impose very drastic restrictions on the original. It has been claimed, in fact, that if conditions (1) and (2) are not actually inconsistent (in which case the class of sequence defined would be empty), there cannot be any wide range of types that satisfy both requirements and that therefore the application to actual statistical data is seriously restricted.

A sequence satisfying the foregoing two conditions is termed by Mises a Collective, and the purpose of his investigation is to show, if possible, that the fundamental theorems of mathematical probability, viz. the Addition, Multiplication, and Bernoulli Theorems,† all hold for a Collective. It would then be possible to state under what conditions these theorems might be validly applied to the analysis of a statistical system.

In the pursuit of this objective great mathematical difficulties have been experienced. To establish the multiplication theorem a special definition has to be made to cover the case of two Collectives that are mutually disorderly. In effect this is met by the requirement that by no systematic transformation can

† See pp. 49, 51, 58.

the one Collective be transformed into the other. On the other hand, to establish Bernoulli's Theorem Mises requires to apply the Principle of Disorder already referred to, not only to sequences transformed according to some law of position, i.e. according to some specific function of n, but also to all those that can be derived by applying any regular rule to localized qualities in the Collective, e.g. deriving a sequence by choosing the numbers that are two places to the left of each 1.

We need not pursue this topic in greater detail. Serious criticisms have been raised against the validity of the Mises approach by Waismann, Kamke, Reichenbach, Popper, and others. It is contended, for example, that condition (1) is in itself meaningless; that there can be no significance to the convergence property without defining the law of the sequence, and since the essence of the sequence is that it should be lawless except for condition (1), there is an inherent contradiction involved. Actually, of course, Mises's first condition is really a demand on the derived convergency sequence. Similar criticisms have been levelled against the suggestions of Kamke and Reichenbach, in their efforts to escape from the various logical dilemmas aroused. The result is that the scope of the Collective becomes so restricted that the class of illustration included reduces almost to emptiness, it becomes increasingly difficult to find actual illustrations that satisfy the requirements, and the statistical value of the approach is thus seriously impaired. The importance of the subject rests therefore rather on the nature of the logical problems raised than on any adequate bridge that may be built between statistical and mathematical probability.

3. Mathematical determinism

Scientific investigation, when used as a guide to action, is turned in the first instance towards making a prediction; it seeks to state that if certain circumstances remain unchanged, then an event will develop in a particular way. In mathematics this process takes the form of strict logical deduction; in statistical work, on the other hand, the process is essentially one of induction, and for that reason the final statement is accompanied with less assurance than the mathematician's. The difference in outlook between the mathematician and the empiricist is, however, more apparent than real; the former has cast aside his doubts by postulating a given set of circumstances and relying on mathematical logic: the latter, in making his induction, is doubtful whether the 'givenness' can be carried forward. Nevertheless, both scientists arrive at a unique and precise conclusion. It is worth while examining how this can occur, since our examination will bring out the part which probability estimates play in the process.

The Typical Problem of Mathematics

Consider the problem of constructing a plane triangle from the knowledge of two sides and the angle included between them. If this knowledge is exact, the triangle can be uniquely constructed, and all its characteristics, e.g. the length of the remaining side and the angles adjacent to it, are uniquely calculable.

This example can be taken as typical of a mathematical problem: certain data are given and certain unique conclusions follow logically. In addition to the 'given' facts, however, there are always certain tacit assumptions implicit in the discussion —in the above case, the assumptions of Euclidean geometry.

Suppose now that the initial data, for the construction of a plane triangle, are two sides and an angle, which is not the included angle. Then it is well known that in general there is no longer a unique solution to the problem; there are in fact two triangles which satisfy the requirements stated. If we asked whether our conditions 'determined' the triangle, the answer would certainly be No. Suppose, however, that having discovered the existence of the two solutions, we restate the problem in the form: To construct the two triangles which have two given sides and a given angle opposite to one of them. The solution to our problem is now unique and has been converted from an indeterminate problem into a determinate one by couching the statement of the problem in appropriate form.

We take another set of data: suppose that it is required to construct the triangle ABC whose base AB is given and whose angle C is a right angle. There is, of course, no such unique triangle; any one of the triangles whose vertex C lies on the circle whose diameter is AB satisfies the given conditions. Thus, at first sight, the problem is necessarily indeterminate, with an infinity of solutions. If, however, we restate the problem by requiring the locus of the vertices of all triangles satisfying the specified conditions, then the locus is unique, and the problem has a unique, determinate solution.† In this, as in the preceding case, there are assumptions inherent in the analysis: we have, for instance, implied that all the required triangles lie in a plane; if we become aware of this restriction and remove it, we obtain for the locus of vertices not a circle but a *sphere.*

These examples illustrate the general proposition that every problem in geometry which starts from a set of data linked together and worked on by a logical process, leads to a unique result which can be regarded as the consequence of a logical determinism.

Problems in classical mechanics are identical in form with such geometrical problems. Once more we are given certain entities—particles of matter, masses, electric charges, etc.—which correspond to the points and lines of the geometrical problem. In addition there are postulated fields of force or interactions between the particles of matter. A typical problem in mechanics may be posed thus: 'A mass M (which we call the sun) is situated in the neighbourhood of another mass m (called the earth); given that the masses are moving with a known speed and attract each other with a force equal to the inverse square of the distance, what follows as regards their paths?' Here again the problem is in reality one of finding a form of statement which, with the given data assembled in mathematical symbols, leads to an inescapable conclusion.

Consider another example: A particle is projected in any *given* direction with a *given* velocity. Given also that the earth's attraction imposes on it an acceleration g downwards, where will the particle meet the horizontal plane through the point of projection? In these circumstances the solution is logically unique and determinate and is applied for the prediction of physical events. This fact is sometimes referred to as mechanical

† Cf. Abel's dictum: 'On doit donner au problème une forme telle qu'il soit toujours possible de le résoudre.'

instead of logical determinism; but to justify such terminology we should require evidence that what is given and what is accepted as logical necessities are both necessities of natural mechanical processes. For the moment the important fact for us is that the conclusion is unique, and in the circumstances inescapable. If instead the particle is projected in *any* direction with a given velocity, it is not difficult to prove that there is no unique solution to our problem. On the other hand, if we require the *maximum range* described by the particle, then once again the solution is unique and determinate.

In classical mechanics, as we have described it, every problem can be posed in such a way that, with the given data and the principles for their combination, its solution is unique and precise, and no indeterminism need arise: the essence of the procedure is deterministic. Now there are two classes of investigation in which this procedure appears to be unsatisfactory, and both arise from the application of the classical method to problems of prediction. As we have seen, the process of prediction is itself necessarily a logical one: if we have appropriately phrased our problem in the light of the data and adopted the correct. physical guiding principles, to obtain anything but a unique solution is, in physical science, tantamount to a failure of science. We therefore ask in what respects may our assumptions and principles be invalid; in so far as they relate to the question of prediction.

The Two Classes of Investigation

Let us examine the two classes of investigation referred to above: both result from the problem of deciding what may be considered as 'given' in the process of prediction in Nature. For his own purposes, the mathematician may assume any set of mutually consistent hypotheses; but in order to satisfy the physicist, these must represent what is actually found in Nature. Thus, in our example of the projected particle, we assumed that the particle is projected with a given velocity in a given direction. The particle may be given, but in practice it is not a mathematical 'point' but a physical 'piece of matter' having size, shape, and weight. Again, the given velocity of projection is, for physical purposes, the velocity *as actually* *measured*; and an elementary knowledge of experimental processes tells us that it is impossible to say precisely what that is. As far as the experimenter's knowledge goes, it may be anything between certain narrow limits, and a variation of even a small amount in the velocity may make a considerable difference in the range of the particle. The mathematician, too, may make a tacit assumption that the particle is projected *in vacuo*: the physicist, who knows better, expects the neglected resistance of the air to make a considerable difference to the range.

There are innumerable other factors, which we need not describe, that cause the actual problem to differ from the mathematical one. Even the final verification to test the mathematical prediction of the range is subject to the same sort of imprecision as the measured 'length' of the desk (p. 16). What does this imply? It means that in assuming a series of initial factors as 'given', the mathematician has followed a mathematically determinate scheme, and has thus tacitly supposed that all the interconnexions of his abstract isolated problem with the rest of the universe can be legitimately ignored. If he proposes to apply such a process to the real world, every one of the so-called 'given' elements in his problem must be introduced not in the form of a discrete quantity, but as one which may vary within a certain band of values, determined for him by the experimenter. The process of prediction can still be carried through and the answer obtained is unique; but it has to be couched, not in the form, 'the resulting range is precisely so much', but in the form, 'the range must lie within a certain band of variation'. We must realize that, in making a prediction, the mathematician endeavours to anticipate the measurement that will actually be found, and that he is concerned only with such measurements: he never discusses the question, *qua* mathematician, whether the process from which these measures emerge is itself determinate apart from this. A prediction, let us repeat, is an attempt to anticipate measurement; and to that extent only is it an attempt to anticipate process.

It will be recognized that the above description of the mathematically determinist process in physics always involves an indeterminacy in a certain special sense: it arises from the gap between the actual process interrelated in Nature and the partial measures of isolated phenomena obtained by the experimenters. It is bound up with the fact that science never studies Nature as a whole but in fragments, tacitly assuming that ideal apparatus can be designed that will be unaffected by the process studied, and that processes can be discovered that are unaffected by the apparatus used. To ignore these inescapable interconnexions, implying that with greater refinement in apparatus and experimental technique the mathematical hypotheses could be made to approximate to any degree of closeness to the physical process, is to be guilty of a methodological fallacy.

Thus the first type of indeterminacy has usually been ascribed to experimental error, the cause of the error being assigned to the so-called 'laws of chance'. Whatever those laws might be, the real implication was that the universe was 'governed' by mechanical laws plus laws of chance; and that if only the latter could be fully elucidated, the mathematician's predictions could be made to coincide absolutely with the experimenter's measurements. It is worth examining in detail why such a coincidence could never occur. Consider this typical illustration. A measuring apparatus has on it a measuring scale subdivided by fine lines: the measuring process consists in fitting a mark between two such subdivisions. Thus in every measurement there is implicit an actual experimental uncertainty, and in an involved experiment, into which many such measurements may enter, the total extent of such uncertainty may be large.

The second class of indeterminacy does not differ fundamentally from the first; the range of experimental uncertainty may be much less important in magnitude but of much deeper physical interest. In our example of the projected particle we have seen that neither the initial position nor the initial velocity can be exactly specified. When, however, the particle is one of sub-atomic nature (e.g. an electron) the statement of the initial conditions presents a special kind of difficulty. To find its position and speed, it would have to be examined, say, through a powerful microscope, and if it is to be visible it must emit at least a quantum of light-energy. But this emission will be accompanied by a rebound on the part of the electron, so that the act of seeing it and measuring its position and speed can only occur physically when its speed and position are in process of change and not otherwise. Here the process of measurement, itself part of Nature, is intimately bound up with and involved in the actual process studied. At the moment it does not appear to be possible to isolate one from the other by any extension of normal scientific method. From a study of the theory and practice of such processes it is found that the product of the uncertainties in the experimenter's measurement of position, and of velocity, is of the magnitude of Planck's constant, a certain well-known number. Thus, the physical limitation involved in the attempt to specify the 'given' conditions for sub-atomic particles leads us to the conclusion that both the initial position and the initial speed cannot be independently determined to any prescribed degree of accuracy, even if the numerous factors already involved in the first class of problem were not present.

Let us emphasize once more the distinction between the two classes of problem. In the first class, despite the uncertainties which arise from the entanglement of the abstracted problem with the rest of the universe, the mathematical logic of the abstract process can still be carried through; in the second class the mathematician who has exposed one of the forms of entanglement is faced with the fact that if he attempts to allow for it initially, the mathematical logic he intended to use no longer avails him. Two quantities which, for the purposes of his logic, should be initially independent, are shown to be interlocked. Accordingly, he is now faced with a new class of problem: given that the initial speed and position are interrelated in the manner described, what are the guiding processes to be assumed for such a group of entities, in order that a unique answer may be obtained, and what will be the general nature of that answer? It must be realized that we are still dealing with a question of mathematical determinism; and although we may find as a result of such an investigation that our prediction asserts that after a certain interval of time the electrified 'particle' may be anywhere within a certain region, this does not vitiate the fact that the process is still deterministic; the problem has only to be correctly stated. The mathematical process determines uniquely for us what can be derived from the given assumed circumstances. The so-called 'uncertainty' resides simply in the physical specification of the assumptions.

A stream of electrons presumably in parallel motion, when striking a film, distribute themselves in the form of concentric rings. The fact that this phenomenon may be described mathematically as 'the probability of an electron falling at any distance from the centre of the film is some kind of function of position', implies not that there is a physical indeterminateness in the fate of any individual electron, but that, in the circumstances, the probability distribution describes the behaviour for the *stream* of electrons. If, therefore, we desire to restate the deterministic conclusions concerning the group-distribution in terms of the behaviour of an individual electron, we can only do this by describing its behaviour in terms of probability. This does not imply an uncertainty in its intrinsic behaviour, but a lack of detailed knowledge for solving the new problem.

CHAPTER III

THE THEORY OF ARRANGEMENTS

SINCE the mathematical theory of probability treats of the relative frequency with which certain groups of objects may be conceived as arranged within a population, one type of problem which we have to consider, preparatory to the main investigation, is concerned with the number of ways in which various sub-groups may be formed or partitioned from the members of a larger group. Many of the theorems arising from this problem are of an elementary nature and to these the present chapter is devoted.

In dealing with objects in groups we are led to consider two kinds of arrangement, according as the order of the objects in the groups is or is not taken into account.

DEFINITION. *The number of different ways in which n objects can be arranged in groups of r, regard being had to the order of arrangement, is called the number of r-permutations of the n objects.*

Evidently, two permutations are identical when they contain the same objects arranged in the same order.

If the n given objects are all different, the number of r-permutations is denoted by the symbol ${}^nP_r$.

To find the number of r-permutations of n different objects

To form any one arrangement we may select any one of the objects to be the first in the arrangement; such a selection can be made in n ways. The second object in our arrangement may be any one of the remaining $n-1$; thus there are $n(n-1)$ ways of arranging the first two objects. Similarly, the selection of the first three objects can be made in $n(n-1)(n-2)$ ways. Thus, in general, we can select r objects in $n(n-1)(n-2)...(n-r+1)$ ways; and therefore

$$ {}^nP_r = n(n-1)(n-2)...(n-r+1). $$

COROLLARY. *The number of n-permutations of n different objects is*

$$ {}^nP_n = n(n-1)(n-2)...3.2.1. $$

The product $n(n-1)(n-2)...3.2.1$ is denoted by the symbol

$n!$, called 'factorial n'. To obtain consistency in our notation, we make the convention that the symbols 1! and 0! are to be interpreted as being equal to unity.

Ex. 1. How many different numbers can be formed by using four out of the nine digits 1, 2, 3,..., 9?

The required number is ${}^9P_4 = 9.8.7.6 = 3{,}024$.

Ex. 2. How many different numbers, each of four digits, can be formed from the ten digits 0, 1, 2,..., 9?

The total number of 4-permutations of the digits is ${}^{10}P_4$, and from this we must deduct the number of permutations in which 0 occupies the first place, that is, ${}^9P_3$. Hence the required number is

$${}^{10}P_4 - {}^9P_3 = 4{,}536.$$

Ex. 3. Show that the number of ways in which n books can be arranged on a shelf so that two particular books are not together is $(n-2)(n-1)!$.

To find the number of permutations of n objects which are not all different

Let the n objects be represented by letters, and suppose that p of them are a's, q of them b's, r of them c's, and so on.

If for a moment we suppose that the p letters a are changed into letters which are different from each other and from the rest, then by changing only the arrangement of these new letters, we should have, instead of one permutation, $p!$ different permutations.

Hence, if P is the required number of permutations, the number of permutations now obtained is $Pp!$.

Similarly, if we suppose that the b's are changed into q letters different from each other and from the rest, the number of permutations is now $Pp!q!$. Proceeding in this manner, we see that if all the letters are changed so that no two are alike, the total number of permutations is $Pp!q!r!\ldots$.

But in this case it is clear that the total number of permutations is $n!$. Hence $Pp!q!r!\ldots = n!$, so that

$$P = \frac{n!}{p!q!r!\ldots}.$$

This result is, apparently, due to Montmort (1708).

Ex. 1. The number of permutations of all the letters of the word *mississippi* is $\dfrac{11!}{4!4!2!} = 34{,}650$.

Ex. 2. Find the number of r-permutations of n objects, when each can be repeated any number of times.

Any one of the n objects can be selected first, and any one of the objects is still available for selection; and so on. Hence the required number is

$$n\,.\,n\,.\,n\ldots = n^r.$$

Ex. 3. Show that the number of permutations of n objects all together, in which r specified objects are to be in an assigned order, is $n!/r!$.

Ex. 4. Prove that ${}^{n+1}P_r = {}^nP_r + r\,{}^nP_{r-1}$.

Definition. *The number of different ways in which n objects can be separated into groups of r, irrespective of the order of arrangement, is called the number of r-combinations of the n objects.*

When the objects are all different, the number of r-combinations is denoted by ${}^nC_r$.

To find the number of r-combinations of n different objects.

It is clear that every such combination would give rise to $r!$ permutations, if the order of the objects were altered in all possible ways. Hence we have

$$r!\,{}^nC_r = {}^nP_r.$$

The same result may be obtained otherwise, as follows: Consider those r-combinations which contain a particular object; evidently the number of such combinations is ${}^{n-1}C_{r-1}$. Thus, in the total number of r-combinations every object occurs ${}^{n-1}C_{r-1}$ times, and therefore the total number of objects included is $n\,{}^{n-1}C_{r-1}$. But since r objects occur in each combination, the total number must also be $r\,{}^nC_r$. We thus derive the relation

$$r\,{}^nC_r = n\,{}^{n-1}C_{r-1}.$$

This holds for all the values of n and r. Changing n into $n-1$ and r into $r-1$, we have in succession

$$(r-1)\,{}^{n-1}C_{r-1} = (n-1)\,{}^{n-2}C_{r-2},$$
$$(r-2)\,{}^{n-2}C_{r-2} = (n-2)\,{}^{n-3}C_{r-3},$$

$${}^{n-r+1}C_1 = n-r+1.$$

Multiplying together corresponding members of these equations and cancelling the common factors, we obtain

$${}^nC_r = n(n-1)(n-2)\ldots(n-r+1)/r!.$$

Note that ${}^nC_r$ may be written as $n!/r!(n-r)!$.

Corollary 1. *The number of r-combinations of n different objects is equal to the number of $(n-r)$-combinations of the n objects.*

For ${}^nC_{n-r} = n!/(n-r)!(n-n+r)! = n!/r!(n-r)! = {}^nC_r$.

Corollary 2. $^nC_r + {^nC_{r-1}} = {^{n+1}C_r}$.

We have

$$^nC_r + {^nC_{r-1}} = \frac{n(n-1)\dots(n-r+1)}{r!} + \frac{n(n-1)\dots(n-r+2)}{(r-1)!}$$

$$= \frac{n(n-1)\dots(n-r+2)}{r!}(n-r+1+r)$$

$$= \frac{(n+1)n(n-1)\dots(n-r+2)}{r!} = {^{n+1}C_r}.$$

We leave as an exercise to the reader the proof of these results from first principles.

Ex. 1. Find the number of diagonals of a polygon of n sides.

The number is

$$^nC_2 - n = \tfrac{1}{2}n(n-1) - n = \tfrac{1}{2}n(n-3).$$

Ex. 2. In how many ways can a committee of 6 be formed from a party of 5 ladies and 8 gentlemen, if the committee is to contain 2 ladies ?

The number of ways of choosing the ladies is $^5C_2$; the number of ways of choosing the gentlemen is $^8C_4$. Thus the number of possible ways is

$$^5C_2 \times {^8C_4} = \frac{5.4}{1.2} \cdot \frac{8.7.6.5}{1.2.3.4} = 700.$$

Ex. 3. If the committee is to contain at most 2 ladies, then the number of possible selections is

$$^5C_2 \times {^8C_4} + {^5C_1} \times {^8C_5} + {^8C_6} = 700 + 280 + 28 = 1{,}008.$$

Ex. 4. Show that, in the n-combinations of $2n$ different objects, the number of combinations in which a particular object occurs is equal to the number in which it does not occur.

Ex. 5. Given n points in a plane such that no two of the lines joining pairs of points are parallel and no three are concurrent save those which pass through one of the given points, in how many points do the lines intersect ?

For the further discussion of problems of arrangement a number of preliminary theorems are required.

Use of Stirling's Theorem

From the above examples it will be noted that the calculation of ^nP_r and ^nC_r, when n and r are large, may be a tedious if not a difficult process. For the purpose of approximate evaluation, it is often convenient to replace the factorial expressions which occur by other expressions to which they tend *asymptotically*.

A formula due to Stirling (1718), which we shall establish later (p. 65), tells us that

$$n! = \sqrt{(2\pi n)}n^n e^{-n}\left(1+\frac{1}{12n}...\right),$$

where $$e = 1+\frac{1}{1!}+\frac{1}{2!}... = 2\cdot71828....$$

Thus the relative error involved in taking only the first term in the above formula is about $\frac{1}{12n}$, i.e. $8/n$ per cent., approximately.

Ex. We have $5! = 120$, while the Stirling formula gives $5! = 118\cdot1$. Again, $10! = 3{,}628{,}800$; Stirling's formula gives 3,598,699.

The Binomial Theorem

Suppose that we are given n letters $a_1, a_2,..., a_n$, and that we wish to evaluate the product

$$(1+a_1)(1+a_2)...(1+a_n).$$

The first term in the expanded form of this product, in which none of the letters occurs, is 1; the next term, in which each letter occurs once, is the sum of all the letters, denoted by $\sum a_1$; the next term consists of the sum of the products of all the letters taken two at a time, denoted by $\sum a_1 a_2$; and so on. The final term is simply the product of the n letters altogether. Thus we have

$$(1+a_1)(1+a_2)...(1+a_n)$$
$$= 1+\sum a_1+\sum a_1 a_2+\sum a_1 a_2 a_3+...+a_1 a_2...a_n.$$

Now suppose that we write $a_1 = a_2 = ... = a_n = x$; the product becomes $(1+x)^n$. The term $\sum a_1$ is evidently $x\,{}^nC_1$, the term $\sum a_1 a_2$ is $x^2\,{}^nC_2$, and so on. Hence, substituting these results, we obtain

$$(1+x)^n = 1+{}^nC_1 x+{}^nC_2 x^2+{}^nC_3 x^3+...+x^n.$$

This expansion is known as *the binomial theorem for a positive exponent n.*

The Binomial Coefficients

We write the binomial expansion in the form

$$(1+x)^n = c_0+c_1 x+c_2 x^2+...+c_r x^r+...+c_n x^n. \tag{1}$$

The coefficient c_r is equal to ${}^nC_r = {}^nC_{n-r}$, by Corollary 1 (p. 43). Thus $c_r = c_{n-r}$; that is, the coefficient of x^r is equal to the coefficient of x^{n-r}.

Putting $x = 1$ in the identity (1) we obtain

$$c_0+c_1+c_2+...+c_n = 2^n.$$

Putting $x = -1$, we have

$$c_0-c_1+c_2-c_3+...+(-1)^n c_n = 0.$$

From these results it follows that

$$c_0+c_2+c_4+... = c_1+c_3+c_5+... = 2^{n-1}.$$

Now put $x = i$, where $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, and so on. Thus

$$c_0+ic_1-c_2-ic_3+c_4+\ldots = (1+i)^n.$$

Putting $x = -i$, we have

$$c_0-ic_1-c_2+ic_3+c_4-\ldots = (1-i)^n.$$

By addition we obtain

$$c_0-c_2+c_4-c_6+\ldots = \tfrac{1}{2}\{(1+i)^n+(1-i)^n\}.$$

By subtraction,

$$c_1-c_3+c_5-\ldots = \frac{1}{2i}\{(1+i)^n-(1-i)^n\}.$$

Ex. 1. By considering the product of $(1+x)^n$ and $(1+1/x)^n$, show that

$$c_0^2+c_1^2+c_2^2+\ldots+c_n^2 = 2n!/(n!)^2.$$

Ex. 2. Find the value of

$$c_0^2-c_1^2+c_2^2-\ldots.$$

Ex. 3. Prove that

$$c_1+2c_2+3c_3+\ldots+nc_n = n2^{n-1}.$$

Greatest Term in the Expansion

In the expansion of $(1+x)^n$, where n is a positive integer, and x is positive, the ratio of the $(r+1)$th term to the rth is evidently

$$\frac{n(n-1)\ldots(n-r+1)}{r!}\cdot\frac{(r-1)!}{n(n-1)\ldots(n-r+2)}x = \frac{n-r+1}{r}x.$$

This ratio can be written as $\left(\frac{n+1}{r}-1\right)x$, and since $\frac{n+1}{r}$ decreases as r increases, the ratio itself decreases as r increases. If the ratio is less than 1 for any value of r, the $(r+1)$th term will be *less* than the rth. Hence, in order that the rth term should be the greatest we must have

$$\frac{n-r+1}{r}x \leqslant 1 \quad\text{and}\quad \frac{n-r+2}{r-1}x \geqslant 1.$$

Thus r satisfies the inequalities

$$r \geqslant \frac{(n+1)x}{x+1}, \qquad r \leqslant \frac{(n+1)x}{x+1}+1.$$

When $r = \frac{(n+1)x}{x+1}$, we have $\frac{n-r+1}{r}x = 1$; in this case there is no one greatest term in the expansion, but the rth and $(r+1)$th terms are equal, and are greater than any of the other terms.

If x is negative, the terms of the expansion alternate in sign, but the method used above still avails to determine the *numerically* greatest term in the expansion.

Ex. Find the greatest term in the expansion of $(1+x)^{10}$, when $x = \frac{2}{3}$.

The Multinomial Theorem

If n is a positive integer, the expression $(x_1+x_2+\ldots+x_m)^n$ may be expanded in a form analogous to that obtained for $(1+x)^n$.

Thus, to distribute the product of n factors

$$(x_1+x_2+\ldots+x_m)(x_1+x_2+\ldots+x_m)\ldots(x_1+x_2+\ldots+x_m)$$

we have to find the coefficient of any given term, for example

$$x_1^{\alpha_1}\, x_2^{\alpha_2} \ldots x_m^{\alpha_m},$$

where $\alpha_1+\alpha_2+\ldots+\alpha_m = n$.

Evidently, the number of times that this particular term arises in the product is the number of n-permutations of the m letters, in which α_1 are alike, α_2 are alike . . ., and so on. Hence by the theorem given above (p. 42) the coefficient of the given term is $\dfrac{n!}{\alpha_1!\,\alpha_2!\ldots\alpha_m!}$.

Thus, finally, we obtain

$$(x_1+x_2+\ldots+x_m)^n = \sum \frac{n!}{\alpha_1!\,\alpha_2!\ldots\alpha_m!}\, x_1^{\alpha_1}\, x_2^{\alpha_2} \ldots x_m^{\alpha_m},$$

where $\alpha_1, \alpha_2,\ldots, \alpha_m$ take all positive integral values for which

$$\alpha_1+\alpha_2+\ldots+\alpha_m = n.$$

This result is the *multinomial theorem* for a positive integral index.

The Binomial Series

If n is not a positive integer, the series

$$1+nx+\frac{n(n-1)}{2!}x^2+\frac{n(n-1)(n-2)}{3!}x^3+\ldots$$

does not terminate; we may show that it converges for all values of x which are numerically less than unity. When n is a negative integer the sum of the series, for such values of x, is equal to $(1+x)^n$; and when n is a rational number the sum is equal to the *principal value* of $(1+x)^n$, i.e. the real positive value of this expression.

Thus, if $n = -m$, where m is a positive integer, then

$$(1+x)^{-m} = 1-mx+\frac{m(m+1)}{2!}x^2-\frac{m(m+1)(m+2)}{3!}x^3+\ldots,$$

provided $x < 1$. In particular, we have

$$(1+x)^{-1} = 1-x+x^2-\ldots,$$

$$(1-x)^{-1} = 1+x+x^2+\ldots,$$

$$(1+x)^{-2} = 1-2x+3x^3-\ldots.$$

To find ${}^nH_r$. The number ${}^nH_r$ of homogeneous products of r letters which can be formed from n given letters may be found by a method which will be employed extensively later. Suppose that thc letters are a, b, c, ... If we form the product

$$(1+ax+a^2x^2+...+a^rx^r)(1+bx+b^2x^2+...+b^rx^r)\times$$
$$\times(1+cx+c^2x^2+...+c^rx^r)...$$

it is at once seen that the sum of the required homogeneous products is the coefficient of x^r in this product. Hence the number of such products is the coefficient of x^r in the product

$$(1+x+x^2+...+x^r)(1+x+x^2+...+x^r)...$$

consisting of n identical factors.

If we suppose that $x < 1$, this is the coefficient of x^r in the expansion of $(1-x)^{-n}$. Thus, by the previous result, we have

$${}^nH_r = \frac{n(n+1)(n+2)...(n+r-1)}{r!} = {}^{n+r-1}C_r.$$

CHAPTER IV

ELEMENTARY THEOREMS ON MATHEMATICAL PROBABILITY

We begin with a restatement of our definition in simplified form:

If there is a class of N letters containing n letters a, then the probability of a letter, specified as belonging to the class N, being a letter a is n/N.

By 'probability' in this chapter it is understood we mean mathematical probability.

Suppose, for example, that we have a group of symbols which are separable into the numbers 1, 2,..., 9, the letters a, b, c, and the letters α, β, γ. A particular symbol is defined as being a member of the whole class. We may then state, on the definition, that the probability that the symbol is a number is $\dfrac{9}{9+3+3} = \dfrac{3}{5}$; the probability that it is a Roman letter is $\dfrac{3}{15} = \dfrac{1}{5}$; and the probability that it is a Greek letter is $\dfrac{3}{15} = \dfrac{1}{5}$.

It should be noticed that the probability that the symbol is a letter and not a number is $\dfrac{3+3}{15} = \dfrac{3}{15}+\dfrac{3}{15} = \dfrac{1}{5}+\dfrac{1}{5} = \dfrac{2}{5}$. Thus the probability that the symbol, defined as a member of the whole class, should be a member of the class consisting of the two subclasses of letters, is the *sum* of the probabilities that it is a member of each of the two subclasses. This result is an illustration of the following general theorem.

Theorem. *An object is defined as belonging to a class of N objects which contains the subclasses of objects a_1, a_2, in number n_1, n_2, respectively, having no members in common. Then if'the probabilities that the object belongs to the subclasses a_1, a_2 be separately p_1 and p_2, the probability that it belongs to the combined group of objects a_1+a_2 is p_1+p_2.*

The proof of this theorem follows at once from the definition. Evidently the result may be extended step by step to give the probability that an object of the class N should belong to the group $a_1+a_2+a_3$, or the group $a_1+a_2+a_3+a_4$; and so on.

Ex. Suppose that we are given a book of N pages such that n_1 of the pages each contain one printer's error, n_2 contain two such errors,..., and generally, n_r contain r errors. Then

the probability that a page has r errors is n_r/N;

the probability that a page has at least r errors is $(n_r+n_{r+1}+...)/N$;

the probability that a page has not less than r and not more than s errors is $(n_r+n_{r+1}+...+n_s)/N$.

For it is clear that if a page has, say, r errors, it cannot have s errors, where r and s are unequal; so that the classes of pages so defined have no members in common, and our theorem can be applied.

It is obvious from our definition that mathematical probability is a number lying between 0 and 1 and that, since it is the ratio of two integers, it must be a proper fraction. We shall have occasion later to extend the definition.

When the probability p is equal to unity, its maximum value is attained; in such a case the class to which the object belongs is identical in extent with the subclass. It is desirable to avoid referring to the case $p = 1$ as 'certainty' for this would seem to imply a psychological state to which our numbers have not necessarily any direct relevance. (One may be certain of the truth of a falsehood.) Similarly, the case $p = 0$ is frequently referred to as representing 'falsehood', and to this the same criticism applies; in point of fact, $p = 0$ is excluded from our consideration, for such a value of p would imply that the subclass is not a member of the whole class.

Mathematical Expectation

Let the letters a_1, a_2,... denote particular classes of events, with which are associated numbers $M_1, M_2, \ldots$. For example, the events might be the actual processes of measuring some object, and M_1, $M_2, \ldots$ the magnitudes obtained. Then the probability of occurrence of the event is also the probability of occurrence of the magnitude.

If p_1 is the probability that the event a_1 will produce a magnitude M_1, then its *mathematical expectation* is defined as $p_1 M_1$.

Thus, a person tosses a coin; if it turns up heads he is to receive a shilling—otherwise he receives nothing. Then the probability of winning a shilling is $\frac{1}{2}$ and the expectation is sixpence.

More generally, in the case of n independent events, for which the probabilities that the events will produce magnitudes M_1, M_2,..., M_n are respectively p_1, p_2,..., p_n, the expectation E

associated with some unspecified event of the set is

$$E = \sum_{r=1}^{n} p_r M_r.$$

Ex. If n measurements, all equally probable, are made of the same length, show that their mathematical expectation is the average value.

THEOREM. *If p is the probability that a member of a class is also a member of a given subclass, then $1-p$ is the probability that it is not a member of that subclass.*

For if the class N can be divided into subclasses having n and $N-n$ members respectively, and $p = n/N$, then the probability that an object is not a member of the class (n) is the probability that it is a member of the class $(N-n)$. This probability is $(N-n)/N = 1-p$, which proves the proposition. For convenience we write $q = 1-p$. If this relation is written in the form $p+q = 1$, it is equivalent to the assertion that 'it is true that an object is either a member of a particular subclass or of the class of remaining objects'.

Ex. 1. The probability that a coin falls either on its head or its tail, given that it falls flat, is 1. If the probability that it falls on its head is $\frac{1}{2}$, then the probability that it falls on its tail is also $\frac{1}{2}$. Thus, the probability that it falls on its head $= 1-$(the probability that it does not).

Ex. 2. In the example (p. 50), the probability that a page has not more than r errors is $1-(n_r+n_{r+1}+...)/N$. The probability that it has no errors is $1-(n_1+n_2+...)/N$.

Ex. 3. Consider two dice each marked with the numbers 1 to 6. It is given that each lies with a face upwards: what is the probability that both faces show fours?

To find the total number of members of the class of pairs of faces, one for each die, we observe that each of the faces of one die may be grouped with each face of the other, thus giving $6\times 6 = 36$ members of the class. There is only one member of the class (4, 4); thus the probability that both faces show fours is $\frac{1}{36}$. We notice that $\frac{1}{36} = \frac{1}{6}\times\frac{1}{6}$, i.e. equals the probability that a face of one die is a four, multiplied by the probability that a face of the other die is a four.

Ex. 4. In a certain examination, 10 of the 30 students receive over, and 20 under, 50 per cent. of the total marks. It is known that two-thirds of the candidates have written their papers in ink and the rest in pencil. An examiner selects a name from the list of 30: what is the probability that the candidate selected wrote his script in pencil and received more than half marks?

These illustrations are typical of the following result:

THEOREM. *If p_1 is the probability that an object belongs to the subclass a_1 of the classes a_1, a_2,..., a_r, and P_1 is the probobility of* *its belonging to the subclass A_1 of the classes $A_1, A_2, \ldots, A_s$ (which are exclusive to $a_1, a_2, \ldots, a_r$), then the probability that it belongs to the combined class $a_1 A_1$ is $p_1 P_1$.*

Let us set out the classes in a scheme, as follows:

First class	$a_1, a_2, a_3, \ldots, a_r$
Number of members . .	$n_1, n_2, n_3, \ldots, n_r$
Probability. . . .	$p_1, p_2, p_3, \ldots, p_r$.
Second class . . .	$A_1, A_2, A_3, \ldots, A_s$
Number of members . .	$N_1, N_2, N_3, \ldots, N_s$
Probability. . . .	$P_1, P_2, P_3, \ldots, P_s$.

Thus $p_1 = \dfrac{n_1}{n_1+n_2+\ldots+n_r}$, and $P_1 = \dfrac{N_1}{N_1+N_2+\ldots+N_s}$.

Combined class	$a_1 A_1, a_1 A_2, \ldots, a_r A_s$
Number of members . .	$n_1 N_1, n_1 N_2, \ldots, n_r N_s$
Total number of members.	$(n_1+n_2+\ldots+n_r)(N_1+N_2+\ldots+N_s)$.

Hence the required probability is

$$\frac{n_1 N_1}{(n_1+n_2+\ldots+n_r)(N_1+N_2+\ldots+N_s)} = p_1 P_1.$$

This theorem, sometimes known as the Multiplication Theorem, may be illustrated geometrically as follows.

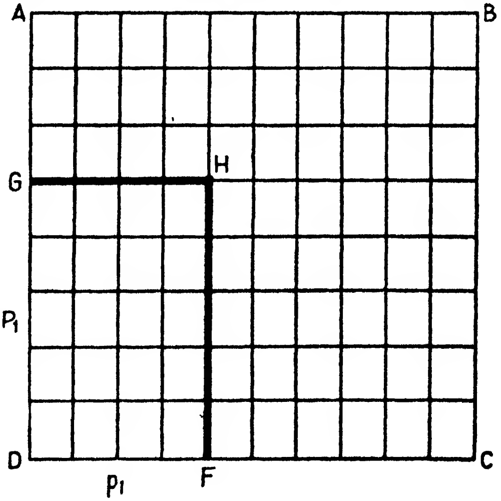


FIG. 1

Let $ABCD$ be a square of unit side, and let DF, DG represent the lengths corresponding to the probabilities p_1 and P_1. Then if DC and DA be subdivided equally, each as many times as there are members of the two classes $a_1, a_2, \ldots, a_r$ and $A_1, A_2, \ldots, A_s$, and rectangles are formed by drawing parallels to DC, DA through the extremities of these subdivisions, the required probability will be the ratio of the number of rectangles within $DGHF$ to the number of rectangles in $ABCD$. Since the area of the latter is unity, the ratio is $p_1 P_1$. The probability of such a combined class is referred to as that of a 'double event'. We note that, if p_1 and p_2 are the successive probabilities of two individual events, the probability of the double event not occurring is $1-p_1p_2$.

Ex. 1. In a certain book of N pages, no page contains more than three errors; n_1 of the pages contain one error, n_2 contain two errors, and n_3 three errors. Two copies of the book are opened at any two given pages. Then the probability that both pages have two errors is n_2^2/N^2; the probability that the total number of errors is 4 is

$$(n_1n_3+n_2^2+n_3n_1)/N^2 = (2n_1n_3+n_2^2)/N^2;$$

the probability that the total number is 5 is

$$(n_2n_3+n_3n_2)/N^2 = 2n_2n_3/N^2;$$

the probability that the total number is 6 is n_3^2/N^2; the probability that the total number is at least 5 is $n_3^2/N^2+2n_2n_3/N^2$; the probability that the total number is not more than 4 is $1-(n_3^2+2n_2n_3)N^2$.

✓Ex. 2. *Tchebycheff's Problem.* Two integers lie within the range 2 to N. What is the probability that they are prime to one another?

Any number, when divided by a suspected prime factor r, may have a remainder $0, 1, \ldots, r-1$; hence the probability that it is divisible by r is $\frac{1}{r}$. Thus the probability that both the integers are divisible by r is $\frac{1}{r^2}$, and, therefore, the probability that both are *not* divisible by r is $1-\frac{1}{r^2}$. It follows that the probability that the two integers have no common prime factor over the whole range is

$$x = \left(1-\frac{1}{2^2}\right)\left(1-\frac{1}{3^2}\right)\ldots\left(1-\frac{1}{p^2}\right),$$

where p is the greatest prime in the given range 2 to N.

If N (and therefore p) is large, we may approximate to x as follows.

We suppose that x is approximately equal to the infinite product

$$\left(1-\frac{1}{2^2}\right)\left(1-\frac{1}{3^2}\right)\left(1-\frac{1}{5^2}\right)\ldots\left(1-\frac{1}{r^2}\right)\ldots,$$

where r is always prime.

Then

$$\frac{1}{x}=\left(1-\frac{1}{2^2}\right)^{-1}\left(1-\frac{1}{3^2}\right)^{-1}\ldots\left(1-\frac{1}{r^2}\right)^{-1}\ldots$$

$$=\left(1+\frac{1}{2^2}+\frac{1}{4^2}\ldots\right)\left(1+\frac{1}{3^2}+\frac{1}{3^4}\ldots\right)\ldots,$$

and since any number is either a prime or a product of primes, it follows on multiplying out that

$$\frac{1}{x}=1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots+\frac{1}{n^2}+\ldots=\frac{\pi^2}{6}.\dagger$$

Hence $x=\dfrac{6}{\pi^2}=\dfrac{3}{5}$, approximately.

Tchebycheff's problem is sometimes stated in the form: to find the probability that the fraction m/n is in its lowest terms, m and n being any two integers.

Note that this process does not give the value of the *probability* (which is necessarily a proper fraction) but only an approximation to it.

In the following example the actual fraction is calculated for the numbers between 2 and 20 and between 2 and 30.

Thus we find that the number of pairs of numbers between 2 and 20, with no common factors, is 108. The total number of pairs is ${}^{19}C_2=171$. Applying this result to find an approximation to π, we have

$$\frac{6}{\pi^2}=\frac{108}{171}=\frac{12}{19}, \text{ giving } \pi^2=9{\cdot}5, \text{ and } \pi=3{\cdot}08.$$

For the range 2 to 30 we find that the number of prime pairs is 248, while the total numbers of pairs is ${}^{29}C_2=406$. These data give

$$\frac{6}{\pi^2}=\frac{248}{406}, \text{ whence } \pi^2=9{\cdot}82, \text{ and } \pi=3{\cdot}13.$$

† See Hobson, *Plane Trigonometry*.

Ex. 3. An interesting application of elementary probability is found in the work of Bunsen and Kirchhoff in connexion with the discovery of the presence of iron in the sun. By comparing the spectra of sunlight and incandescent iron vapour it was found that, to the degree of accuracy given by the instruments, 60 bright lines coincided in the two spectra. Now the average distance between the solar lines in Kirchhoff's map was 2 mm., and coincidence for his instruments implied that a line from the iron vapour must fall within $\frac{1}{2}$ mm. on either side. Thus the probability of casual coincidence for each of the 60 lines was $2.\frac{1}{2}/2 = \frac{1}{2}$. Accordingly, the probability of casual coincidence for all 60 lines was $\frac{1}{2^{60}}$, or one in a million million millions. It should be noted that in this analysis iron is *defined* as that substance which gives the above 60 lines in the spectrum.

Similar considerations with regard to the coincidence of the spectra of solar, lunar, and planetary light can be used to decide the probability that they are all of the same origin.

EXAMPLES ON CHAPTER IV

[In the following examples it is to be assumed that when the phrase 'a coin is tossed' is used, it is implied that the probability of the appearance of a head is $\frac{1}{2}$. See also Chapter V.]

Ex. 1. What is the probability of a penny turning up heads at least once in n throws?

The probability that it turns up tails every time is $\frac{1}{2^n}$. Hence the probability that it shows heads *at least* once is $1-\frac{1}{2^n}$.

Ex. 2. If m coins are tossed and all the heads are removed, and then the remaining coins are tossed and the heads removed, and so on, what is the probability that all the coins will be removed by or before n tossings?

We may imagine all the coins tossed n times; we thus require the probability that each will turn up heads at least once in n tossings. Hence the required probability is $\left(1-\frac{1}{2^n}\right)^m$.

Ex. 3. (Pascal's and Fermat's problem.) Two players, with equal probability of winning a point, agree to play a game for 5 points. If the game must not be drawn, find their respective chances of winning at any given stage of the game.

Ex. 4. An urn contains b black and w white balls. If n balls are extracted together, what is the probability that α of these are white?

The number of ways in which n balls can be extracted is ${}^{b+w}C_n$. The number of sets of n balls which contain α white balls is

$${}^{w}C_{\alpha}.{}^{b}C_{n-\alpha}.$$

Ex. 5. A number is chosen from each of the two sets 1, 2, 3,..., 9; 1, 2, 3,..., 9. Show that the probability that the sum of the numbers should be 10 is $\frac{1}{9}$, and that their sum should be 8 is $\frac{7}{81}$.

Ex. 6. If in selecting a number from the set 1, 2, 3,..., 9, 7 is chosen twice as often as 3, 3 twice as often as 5 and 9, and 5 and 9 twice as often as 1, 2, 4, 6, 8, what is the probability that the sum of two numbers selected will be 10?

Ex. 7. A red card is removed from a pack of 52; 13 cards are then drawn and found to be of the same colour. Show that the odds are 2 to 1 that the colour is black.

Ex. 8. A set consists of n counters. What is the probability that a selected group of these of unspecified number consists of (1) an even number of counters, (2) an odd number of counters?

We have to find the total number of members of the groups that can be formed of 2, 4, 6,... counters for the case (1) and of 1, 3, 5,... for the case (2).

The total number of ways of forming groups of 2, 4, 6,... is respectively ${}^{n}C_2$, ${}^{n}C_4$, ${}^{n}C_6$,... and for forming the groups 1, 3, 5,... is

$${}^{n}C_1,\ {}^{n}C_3,\ {}^{n}C_5,\dots.$$

Thus the number of members of the class of even groups is

$${}^{n}C_2+{}^{n}C_4+\dots = 2^{n-1}-1 \quad \text{(p. 45)}$$

and the number of members of the class of odd groups is

$${}^{n}C_1+{}^{n}C_3+\dots = 2^{n-1},$$

while the total number of members of all classes is 2^n-1. Thus the probability of the selected group being odd is greater than its being even. The difference between the two probabilities decreases as n increases.

Ex. 9. From a pack of 52 cards an even number of cards is drawn. Show that the probability that these consist half of red and half of black is

$$\left\{\frac{52!}{(26!)^2}-1\right\}\Big/(2^{51}-1).$$

The number of ways in which an even number of cards can be drawn is

$${}^{52}C_2+{}^{52}C_4+\dots+{}^{52}C_{52} = 2^{51}-1 \quad \text{(p. 45).}$$

Of these, the number of groups consisting half of red and half of black cards is

$${}^{26}C_1^2+{}^{26}C_2^2+\dots+{}^{26}C_{26}^2 = \frac{52!}{(26!)^2}-1 \quad \text{(p. 46).}$$

Hence the result.

Ex. 10. Using Stirling's theorem, find an approximation to this probability.

Ex. 11. A pack of 52 cards is cut twice, a card drawn and replaced; show that the probability of obtaining aces each time is 1/169.

Ex. 12. A and B stand in a line with 10 other persons. What is the probability that there are 3 persons between A and B? What would be the probability if they stood in a ring?

Ex. 13. Find the probability that a month contains portions of six different weeks.

Ex. 14. Two identical urns contain respectively n and n' balls; the first urn contains a white balls and the second a'. If a ball is extracted from one of the two urns, what is the probability that it is white?

It must be noticed that the extraction of a white ball from the first urn is the result of two circumstances: (1) the choice of this urn from the two identical urns, (2) the extraction of a white ball from this urn, supposing that it has actually been chosen. The probability of (1) is $\frac{1}{2}$, that of (2) is $\frac{a}{n}$. Thus, the probability of extraction of a white ball from the urn is $\frac{1}{2}\frac{a}{n}$; and similarly, the probability of extraction of a white ball from the second urn is $\frac{1}{2}\frac{a'}{n'}$. Hence the required probability, which is the sum of these two probabilities, is $\frac{1}{2}\left(\frac{a}{n}+\frac{a'}{n'}\right)$.

Ex. 15. Each of two bags contains m shillings and n sixpences. If a coin is drawn from each bag, show that the probability that both coins are shillings is greater than that of drawing two shillings from a bag containing all the coins.

Ex. 16. In a card game in which the dealer's last card determines the trump suit, find how many hands must be dealt in order that it is more likely than not that at some stage the dealer will hold all the trumps.

Since the dealer always holds one of the trumps, the probability of any one deal of the required type is $\frac{1}{{}^{51}C_{12}}=\frac{1}{c}$, say, where c is a large number.

The probability of not holding all the trumps is thus $1-\frac{1}{c}$. After x deals, this probability is

$$\left(1-\frac{1}{c}\right)^x=\left(1-\frac{1}{c}\right)^{-c(-x/c)}$$

$$=e^{-x/c}, \text{ approximately.}$$

For an even chance we require $e^{-x/c}=\frac{1}{2}$.

This equation gives $x=c\log 2=10^{11}$, approximately.

CHAPTER V

BERNOULLI'S THEOREM

1. Bernoulli's Theorem and its extensions

In dealing with a class of objects or events, we shall use the term 'population' to describe the original class from which the subclasses are to be formed.

Suppose that we are given a population of ten counters divided into two subclasses which we represent by four black counters b and six white counters w. What is the probability that among three unspecified members of the population just two are members of the subclass w?

We may proceed as follows. The probability that a member of the population is a member of w is $\frac{6}{10} = \frac{3}{5}$; hence the probability that two members, as a group, are members of w, is $\frac{3}{5}\times\frac{3}{5}$. To satisfy our conditions, the third member must not belong to w; thus the probability required would appear to be $\frac{3}{5}\times\frac{3}{5}\times(1-\frac{3}{5})$. But the order in which the three members have been considered as belonging (or not) to the subclass w is not exhausted by this particular process; it could be either the second or the first member which is excluded from w. Thus the total probability is

$$3\times\tfrac{3}{5}\times\tfrac{3}{5}\times(1-\tfrac{3}{5}),\quad\text{or}\quad {}^3C_2\times\tfrac{3}{5}\times\tfrac{3}{5}\times(1-\tfrac{3}{5}) = \tfrac{54}{125}.$$

This simple problem is an illustration of the general result.

Bernoulli's Theorem†

Let a population be divisible into subclasses b and w such that the probability of any member of the population being also a member of w is p. Then, of n objects defined only as members of the population, the probability that r of these are also members of w is ${}^nC_r p^r(1-p)^{n-r}$.

For the probability of r members of the population being members of w is, as we have seen, p^r; the probability that the remaining $n-r$ members are *not* members of w is $(1-p)^{n-r}$. Thus the combined probability of the double event is $p^r(1-p)^{n-r}$. But the r members of the group of n initially considered can be

† In interpreting the probability p in the following theorem, reference should be made to the discussion on *a priori* probability on p. 19.

exhaustively selected in ${}^nC_r$ ways. Then, since the total probability required is the sum of the separate probabilities, it is equal to

$${}^nC_r p^r(1-p)^{n-r}.$$

Ex. 1. Thirteen cards are drawn one by one from an ordinary pack of 52, each card being replaced immediately after drawing: to find the probability that exactly 3 red cards are so obtained.

There are initially 26 red and 26 black cards in the pack, so that the probability p that a card should be red is $\frac{1}{2}$. In our theorem, as applied to the present problem, the group of objects to be considered is in number $n = 13$, and the sub-group is in number $r = 3$. Hence the required probability is ${}^{13}C_3\left(\frac{1}{2}\right)^3\left(1-\frac{1}{2}\right)^{13-3} = \frac{143}{2^{12}} = \frac{143}{4{,}096} = \frac{1}{28}$, approximately.

Note, in contrast, that the probability of finding 3 red cards in a hand of 13, as ordinarily dealt, is ${}^{26}C_3\,{}^{26}C_{10}/{}^{52}C_{13}$.

Ex. 2. What is the number of red cards, in such an extraction, for which the probability is greatest?

We have to find the value of r which makes ${}^{13}C_r\frac{1}{2^r}\left(1-\frac{1}{2}\right)^{13-r}$ have its greatest value.

Evidently this is attained when $r = 6$ or 7.

Ex. 3. What is the probability that no more than three of the cards should be red?

This is the sum of the probabilities that the number of red cards should be 0, 1, 2, or 3.

Ex. 4. Find the probability that the hand should contain at least three red cards.

Ex. 5. What is the probability that, in 13 drawings, with replacement, an ace should be obtained four times?

The original probability that a card should be an ace is $\frac{4}{52} = \frac{1}{13}$.

Hence the required probability is ${}^{13}C_4\frac{1}{13^4}\left(\frac{12}{13}\right)^9 = 0{\cdot}02$, approximately.

It should be noticed that the fact that the four specified cards are to be aces is quite irrelevant to the problem; the same probability would be found for the occurrence of *any* four previously indicated cards.

From Bernoulli's Theorem we at once derive the following:

Theorem. ***If p is the initial probability that a member of a population should belong to a specified subclass, the probability that out of n members not more than r belong to this subclass is***

$${}^nC_0(1-p)^n + {}^nC_1 p(1-p)^{n-1} + \ldots + {}^nC_r p^r(1-p)^{n-r}.$$

With the same hypotheses we have:

THEOREM. *The probability that not less than r members belong to the specified subclass is*

$$^{n}C_{r}p^{r}(1-p)^{n-r}+{}^{n}C_{r+1}p^{r+1}(1-p)^{n-r-1}+\ldots+{}^{n}C_{n}p^{n}.$$

Ex. 1. Out of a population of pennies, of which half lie head up and half tail up, a class of $2n$ members is defined. What is the probability that these show heads in excess or defect of n, by a number t?

Evidently the required probability is

$$^{2n}C_{n-t}p^{n-t}(1-p)^{n+t}+\ldots+{}^{2n}C_{n+t}p^{n+t}(1-p)^{n-t}, \quad \text{where } p = \tfrac{1}{2}.$$

This type of problem is usually stated in the form: A penny is tossed $2n$ times. What is the probability that the deviation from n heads should not exceed t? Note that in attempting to identify these two problems we tacitly assume that the sample of $2n$ tossings is drawn from a larger hypothetical population containing precisely the same number of exposed heads as tails.

Ex. 2. With the same interpretation of the terms, show that, if a penny is tossed n times, the probability of not more than r heads is

$$\frac{1}{2^{n}}({}^{n}C_{0}+{}^{n}C_{1}+\ldots+{}^{n}C_{r}).$$

Applications of Mathematical Probability

It will be observed that the language in which these theorems have been developed and the form in which the examples have been couched have been such as scrupulously to avoid all idea of *experiment*. If we are to restrict our investigations in this way we shall certainly avoid the error of confusing psychological expectation with mathematical probability; but we shall also lose the possibility of applying the theory to actual cases. What we have to discover are the circumstances in which such application is legitimate. It was pointed out previously that the study of psychological probability ought logically to follow in the wake of the mathematical investigation. At this stage, therefore, we propose to examine briefly the restrictions hitherto imposed, and to see if they can be circumvented.

It must be understood, then, that when we say that 'a card is drawn from a pack', we mean in fact that we are to discuss certain properties of an entity defined only as a member of the pack. In the same way, when we say that an individual tosses a penny n times, we mean that n events are under consideration and that each of them may belong to one of two classes, head or tail: that is the defining property of the event. If the result is used in any particular case, the onus of justification is on the user who asserts that this defining property is the *only* relevant one in the circumstances in which he applies the result.

In this connexion we may remark that in certain circumstances it is possible to introduce into the defining properties conditions relating to the mode of selection or arrangement which enable the mathematical treatment to provide an answer which is closer to the facts than the result arrived at on a simple hypothesis. For example, suppose that there are ten counters in a row: five black on the left and five white on the right; if all that is asserted about a counter is that it belongs to this group, then on our definition the probability of its being white is $\frac{1}{2}$. If, however, we assert that an individual has selected a counter, then the fact that individuals more frequently choose with their right hand than with their left, and thus more frequently choose an object to the right of the centre of the group than to the left, will vitiate our original calculations and we must introduce a new factor which takes this human bias into account.

Now suppose it is known that the choice made by an individual justifies our statement that the probabilities of choice of the counters, from left to right, are proportional to the numbers

$$\underset{\text{(black)}}{1,\ 1,\ 1,\ 1,\ 2;} \qquad \underset{\text{(white)}}{3,\ 4,\ 4,\ 2,\ 1.}$$

Then the problem may be recast in the form: Given a set of 20 counters of which 6 are black and 14 are white, the probability of a white counter is $\frac{14}{20}$. Thus, by introducing 'weighting' factors to represent the bias in choice of the counters, we have brought the original problem nearer to actuality. In the mathematical problem, these weighting factors must be supposed given; actually, they are given as a result of previous experiment, so that in such a problem they become known *a priori*.

Ex. A sniper finds that, on the average, he kills once in three shots. He fires three times at an enemy; on the assumption that his *a priori* probability of killing is $\frac{1}{3}$, what is the probability that he kills him?

Here we require the probability that *at least* one of the shots should be a hit. Since $p = \frac{1}{3}$, the required probability is

$${}^3C_1\,\tfrac{1}{3}(\tfrac{2}{3})^2 + {}^3C_2(\tfrac{1}{3})^2(\tfrac{2}{3}) + {}^3C_3(\tfrac{1}{3})^3 = \tfrac{19}{27}.$$

Alternatively, we may proceed as follows. The probability of not killing at the first attempt is $\frac{2}{3}$; thus, the probability of not killing in all three attempts is $(\frac{2}{3})^3 = \frac{8}{27}$. Hence the probability of a hit is

$$1-\tfrac{8}{27} = \tfrac{19}{27}.$$

We note that there is no contradiction between this result and the statement that $p = \frac{1}{3}$, for the probability of killing with one shot *only* is $\frac{1}{3}$: after that the probability increases.

Greatest Value of ${}^nC_r p^r(1-p)^{n-r}$

To determine the value of r for which the Bernoulli probability, $B(n,r) \equiv {}^nC_r p^r(1-p)^{n-r}$, has its greatest value, where r is an integer, we cannot legitimately discuss the variation of $B(n,r)$ as a continuous function of r; we are not seeking for a *maximum* but for a greatest value (if it exists) in the range $0 \leqslant r \leqslant n$.

Accordingly, we require to find the value of r such that

$$B(n,r-1) \leqslant B(n,r) \geqslant B(n,r+1),$$

i.e. such that

$${}^nC_{r-1}p^{r-1}(1-p)^{n-r+1} \leqslant {}^nC_r p^r(1-p)^{n-r} \geqslant {}^nC_{r+1}p^{r+1}(1-p)^{n-r-1}.$$

Cancelling out positive factors in common it follows that

$$np+p \geqslant r \geqslant np-(1-p).$$

Since p and $1-p$ are fractions, we thus require that r should be equal to np, if this number is integral, or to the smallest integer greater than np, if np is not integral. We thus obtain the following result.

The greatest value of Bernoulli's probability $B(n,r)$ is obtained by taking r to be np, or the least integer greater than np if np is not integral.

Ex. How many aces are 'most likely' to be found in 13 successive drawings, followed by replacements, from a pack of 52 cards?

First Generalization of Bernoulli's Theorem

Let a population be divisible into subclasses w_1, w_2,..., w_s, the probabilities attached to the subclasses being p_1, p_2,..., p_s. Then, the probability that a group of n members of the population, otherwise unspecified, should contain r_1 members of w_1, r_2 of w_2,..., and r_s of w_s, is

$$\frac{n!}{r_1!\,r_2!\ldots r_s!}p_1^{r_1}p_2^{r_2}\ldots p_s^{r_s},$$

where $$r_1+r_2+\ldots+r_s = n.$$

For, the probability of r_1 members of the population being members of w_1 is $p_1^{r_1}$; of r_2 members belonging to w_2 is $p_2^{r_2}$, and so on. Thus the probability of the combined event is $p_1^{r_1} p_2^{r_2} \ldots p_s^{r_s}$. But the required probability is the sum of the ways in which this combination can be formed subject to the condition that the total number of members is n. Hence we have to multiply $p_1^{r_1} p_2^{r_2} \ldots p_s^{r_s}$ by the number of ways in which a term of this type can arise by n-combinations of all such p's; such a number is identical with the coefficient of $p_1^{r_1} p_2^{r_2} \ldots p_s^{r_s}$ in the expansion of $(p_1+p_2+\ldots+p_s)^n$ (p. 47), whence the result.

The original Bernoulli Theorem follows from this by putting $p_1 = p$, $p_2 = 1-p$, $r_1 = r$, $r_2 = n-r$.

Ex. A pack of 10 cards consists of 3 aces, 2 kings, 2 queens, and 3 jacks. All that is known of them is that on eight successive occasions the cards have been shuffled and the top card each time exposed. It is required to find the probability that an ace will have been top card on two occasions, a queen on three occasions, and a jack on three occasions.

If we denote by w_1, w_2, w_3, and w_4 the respective subclasses defined by the aces, kings, queens, and jacks, then in the previous notation we have

$$n = 8, \quad r_1 = 2, \quad r_2 = 0, \quad r_3 = 3, \quad r_4 = 3,$$
$$p_1 = \tfrac{3}{10}, \quad p_2 = \tfrac{2}{10}, \quad p_3 = \tfrac{2}{10}, \quad p_4 = \tfrac{3}{10}.$$

Hence the required probability is

$$\frac{8!}{2!\,0!\,3!\,3!}\left(\frac{3}{10}\right)^2\left(\frac{2}{10}\right)^0\left(\frac{2}{10}\right)^3\left(\frac{3}{10}\right)^3 = \frac{8!\,27}{10^8} = \frac{108{,}864}{10^7} = \frac{1}{100},$$

approximately.

Alternative statement of Bernoulli's Theorem. The probability that an event with initial probability p occurs exactly r times in n trials is the rth term in the expansion of $(p+q)^n$ in ascending powers of p, where $q = 1-p$.

It follows that the sum of the probabilities for all values of r, is unity.

Again, the *average value of r* in n trials is

$$\sum_{r=0}^{n} {}^nC_r p^r q^{n-r}.r.$$

Now $$(p+q)^n = \sum_{r=0}^{n} {}^nC_r p^r q^{n-r}.$$

Differentiating this identity with respect to p, and then putting $p+q = 1$, we have

$$n = \sum_{r=0}^{n} {}^nC_r r p^{r-1} q^{n-r},$$

and thus $$np = \sum_{r=0}^{n} {}^nC_r p^r q^{n-r}.r.$$

But this is also an approximation to the most probable value of r. It follows that to this degree of accuracy the *average value* of r is the *most probable value* when all possibilities are taken into consideration.

Case of Probability varying from one Trial to another

It has been assumed throughout the foregoing analysis that the successive stages in the withdrawal of a sample from a population are not accompanied by any change in the probabilities of its subclasses; this is the case when, for example, the population consists of a set of black and white balls, and the ball is replaced after each withdrawal, or where the population is generated by an operation, as in the tossing of a coin. If, however, this is not done, the proportion of black to white balls is altered at each stage of the process, and the initial probability of a black or white ball becomes a function of the number of samples.

Ex. 1. If the probability of failing at the nth trial is $1/(1+n)$, what is the probability of succeeding at least once in the first m trials?

Ex. 2. If the probability of failure at the nth trial is $1/2^n$, find the probability of succeeding at least once in three trials.

Second Generalization of Bernoulli's Theorem

Instead of referring to a population and the probability of its subclass, we may speak of an event and the prebability of its success in one or more trials (corresponding, for instance, to the extraction of one or more white balls from an urn containing black and white balls). Suppose then that we consider n independent events whose probabilities of success are $p_1, p_2 \ldots, p_n$; thus the corresponding probabilities of failure are $q = 1-p_1$, $q_2 = 1-p_2, \ldots, q_n = 1-p_n$. Then the probability of obtaining exactly r successes in the compound event is

$$\sum p_i p_j p_k \ldots q_l q_m \ldots,$$

the summation extending to all products of n different symbols, each containing r p's and $n-r$ q's. It will be noticed that this is the coefficient of x^r in the product

$$(p_1 x+q_1)(p_2 x+q_2)\ldots(p_n x+q_n).$$

Hence,

The probability of obtaining r successes in a compound event, consisting of n independent events, is equal to the coefficient of x^r

in $(p_1 x+q_1)(p_2 x+q_2)...(p_n x+q_n)$, *where* p_s, q_s *are the respective probabilities of success and failure in the sth event.*

Ex. Given three urns, of which the first contains 3 white and 4 black balls, the second contains 2 white and 3 black balls, and the third contains 3 white and 5 black balls, what is the probability of obtaining one white ball in extracting a ball from each urn?

Evidently the required probability is the coefficient of x in

$$(\tfrac{3}{7}x+\tfrac{4}{7})(\tfrac{2}{5}x+\tfrac{3}{5})(\tfrac{3}{8}x+\tfrac{5}{8}), \quad \text{i.e.} \quad \tfrac{121}{280}.$$

2. Bernoulli's Theorem and the normal law

Stirling's Theorem

We have already noted (p. 44) the use that an approximate formula for $n!$ may have in evaluating probabilities. In what follows the use of such approximations is essential.

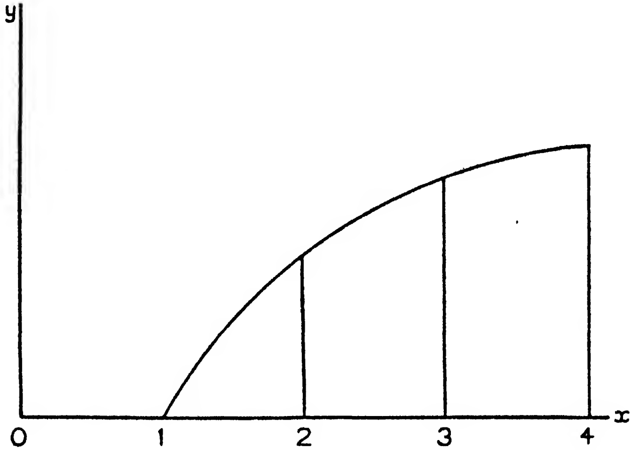


Fig. 2.

We begin by finding an approximation, for large values of n, to

$$\log n! = \log n+\log(n-1)+...+\log 2.$$

Consider the curve representing the function

$$y = \log x.$$

If ordinates be erected at $x = 1, 2,..., n$, then the sum of the trapezia determined by successive pairs of ordinates will be less than the total area between the curve, the nth ordinate, and the x-axis.

Thus

$$\int\limits_1^n \log x\,dx > \tfrac{1}{2}(\log 1+\log 2)+\tfrac{1}{2}(\log 2+\log 3)+\ldots$$
$$\ldots+\tfrac{1}{2}(\log\overline{n-1}+\log n),$$

i.e. $$[x\log x-x]_1^n > \log 2+\log 3+\ldots+\log n-\tfrac{1}{2}\log n,$$

or $$n\log n-n+1 > \log n!-\log n^{\frac{1}{2}},$$

or $$\log n^n e^{-n+1} > \log n!\,n^{-\frac{1}{2}}.$$

Since the logarithms are positive we have

$$n^n e^{-n+1} > n!\,n^{-\frac{1}{2}},$$

or $$n! < n^{n+\frac{1}{2}}e^{-n+1}.$$

It is clear that to obtain a closer approximation to $n!$ we require a more exact estimate of $n!/n^{n+\frac{1}{2}}e^{-n}$.

Write $u_n = \log(n!/n^{n+\frac{1}{2}}e^{-n})$. Then

$$u_{n+1}-u_n = \log\left\{\frac{(n+1)!}{(n+1)^{n+\frac{3}{2}}e^{-n-1}}\cdot\frac{n^{n+\frac{1}{2}}e^{-n}}{n!}\right\}.$$
$$= 1+\log\left(\frac{n}{n+1}\right)^{n+\frac{1}{2}}$$
$$= 1-(n+\tfrac{1}{2})\log\left(1+\frac{1}{n}\right)$$
$$= 1-(n+\tfrac{1}{2})\left\{\frac{1}{n}-\frac{1}{2n^2}+\frac{1}{3n^3}-\frac{1}{4n^4}+\ldots\right\}$$
$$= -\frac{1}{12n^2}+\frac{1}{12n^3}-\frac{3}{40n^4}+\frac{1}{15n^5}-\ldots$$
$$= -\frac{1}{12}\left(\frac{1}{n^2}-\frac{1}{n^3}+\frac{1}{n^4}-\frac{1}{n^5}+\ldots\right), \text{ approximately,}$$
$$= -\frac{1}{12n(n+1)}, \text{ approximately,}$$
$$= \frac{1}{12}\left[\frac{1}{n+1}-\frac{1}{n}\right].$$

Accordingly we may write

$$u_n = A+\frac{1}{12n},$$

where A is an unspecified constant.

Hence we derive

$$\frac{n!}{n^{n+\frac{1}{2}}e^{-n}} = Be^{1/12n},$$

where B is another unspecified constant.

To find its value we have only to use the above approximation for $n!$ for a particular case, say $n = 9$. This gives

$$B = \frac{9!}{9^9 . 3 . e^{-9}}, \quad \text{approximately.}$$

There is, however, a more general method of approach to the evaluation of B. We begin with the well-known formula

$$\sin \pi x = \pi x\left(1-\frac{x^2}{1^2}\right)\left(1-\frac{x^2}{2^2}\right)\left(1-\frac{x^2}{3^2}\right)\ldots.$$

When $x = \frac{1}{2}$, we have

$$1 = \frac{\pi}{2}\left(1-\frac{1}{2^2}\right)\left(1-\frac{1}{4^2}\right)\left(1-\frac{1}{6^2}\right)\ldots,$$

or
$$\frac{2}{\pi} = \frac{1.3}{2^2} \cdot \frac{3.5}{4^2} \cdot \frac{5.7}{6^2} \ldots = \frac{1^2.3^2.5^2.7^2\ldots}{2^2.4^2.6^2\ldots} = \frac{1^2.2^2.3^2.4^2\ldots}{2^4.4^4.6^4\ldots}.$$

Thus
$$\frac{2}{\pi} = \lim_{n\to\infty} \frac{[(2n+1)!]^2}{(2n+1)(n!)^4 2^{4n}}.$$

Inserting our approximation for $n!$, we find that

$$\frac{2}{\pi} = \lim_{n\to\infty} \frac{\left[B(2n+1)^{2n+\frac{3}{2}}\exp(-2n-1)\exp\left\{\frac{1}{12(2n+1)}\right\}\right]^2}{2^{4n}(2n+1)\left[Bn^{n+\frac{1}{2}}\exp(-n)\exp\left(\frac{1}{12n}\right)\right]^4}$$

$$= \lim_{n\to\infty} \frac{1}{B^2}\exp\left\{-4n-2+4n+\frac{1}{6(2n+1)}-\frac{1}{3n}\right\} \cdot \frac{(2n+1)^{4n+2}}{2^{4n}n^{4n+2}}$$

$$= \frac{1}{B^2e^2}\lim_{n\to\infty} 2^2\left(\frac{2n+1}{2n}\right)^{4n+2} = \frac{4}{B^2e^2}\lim_{n\to\infty}\left(1+\frac{1}{2n}\right)^{4n}\left(1+\frac{1}{2n}\right)^2$$

$$= \frac{4}{B^2}.$$

Hence $B = \sqrt{(2\pi)}$ and, finally, we have the approximate formula, for large values of n,

$$n! = \sqrt{(2\pi)}n^{n+\frac{1}{2}}.\exp\left(-n+\frac{1}{12n}\right)$$

$$= \sqrt{(2\pi)}n^{n+\frac{1}{2}}e^{-n}\left[1+\frac{1}{12n}\right], \text{ approximately.}$$

For comparison, it may be remarked that we have already found that

$$n! < n^{n+\frac{1}{2}}e^{-n}.e.$$

Approximate Value of Bernoulli's Probability (case $p = \frac{1}{2}$)

Suppose that we are given a population of coins which shows an equal number of heads and tails; we seek the probability that in a large sample of $2n$ coins there shall be $n+r$ heads and $n-r$ tails, that is, that the excess of the number of heads over the number of tails is $2r$. In this case, the probability p of a head or tail is $\frac{1}{2}$, and Bernoulli's probability gives us the formula

$$\begin{aligned} P &= {}^{2n}C_{n+r}(\tfrac{1}{2})^{n+r}(1-\tfrac{1}{2})^{n-r} \\ &= \frac{(2n)!}{(n+r)!(n-r)!}\,\frac{1}{2^{2n}}. \end{aligned}$$

Using Stirling's formula we write

$$\begin{aligned} (2n)! &= \sqrt{(2\pi.2n)}(2n)^{2n}e^{-2n} = 2^{2n+1}n^{2n+\frac{1}{2}}e^{-2n}\sqrt{\pi}, \\ (n+r)! &= \sqrt{\{2\pi(n+r)\}}(n+r)^{n+r}e^{-n-r}, \\ (n-r)! &= \sqrt{\{2\pi(n-r)\}}(n-r)^{n-r}e^{-n+r}, \end{aligned}$$

so that

$$\begin{aligned} (n+r)!(n-r)! &= 2\pi e^{-2n}(n+r)^{n+r+\frac{1}{2}}(n-r)^{n-r+\frac{1}{2}} \\ &= 2\pi e^{-2n}(n^2-r^2)^{n+\frac{1}{2}}\left(\frac{n+r}{n-r}\right)^r \\ &= 2\pi e^{-2n}n^{2n+1}\left(1-\frac{r^2}{n^2}\right)^{n+\frac{1}{2}}\left(\frac{n+r}{n-r}\right)^r. \end{aligned}$$

Hence

$$\begin{aligned} P &= \frac{(2n)!}{(n+r)!\,(n-r)!}\,\frac{1}{2^{2n}} \\ &= \frac{2^{2n+1}n^{2n+\frac{1}{2}}e^{-2n}}{2\sqrt{\pi}e^{-2n}n^{2n+1}}\left(1-\frac{r^2}{n^2}\right)^{-n-\frac{1}{2}}\left(\frac{n-r}{n+r}\right)^r\frac{1}{2^{2n}} \\ &= \frac{1}{\sqrt{(n\pi)}}\left(1-\frac{r^2}{n^2}\right)^{-n}\left(1-\frac{r}{n}\right)^{r-\frac{1}{2}}\left(1+\frac{r}{n}\right)^{-r-\frac{1}{2}}. \end{aligned}$$

It will appear from the more general investigation on p. 71 that, when r/n is small, the approximate value of P is $\dfrac{1}{\sqrt{(\pi n)}}e^{-r^2/n}$.

Accordingly, $P = \dfrac{1}{\sqrt{(\pi n)}} e^{-r^2/n}$ is the approximate probability that, in a sample of magnitude $2n$, there will be a discrepancy of r heads above (or below) n, provided r/n is small in comparison with unity.

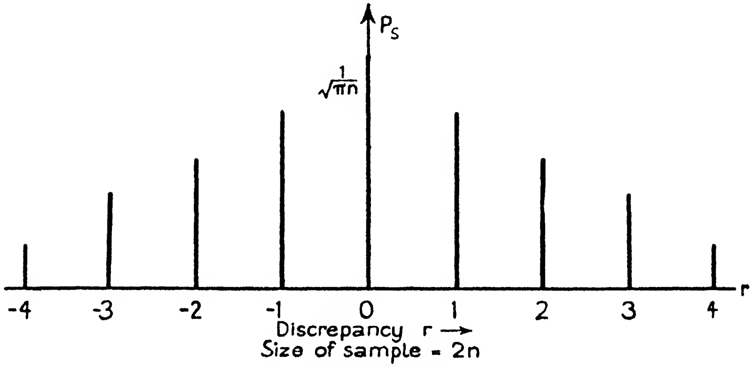


FIG. 3.

In a sample of size $2n$ the probability that the number of heads will lie between $n+s$ and $n-s$ is therefore approximately

$$P_s = \sum_{r=-s}^{r=s} \frac{1}{\sqrt{(\pi n)}} e^{-r^2/n}, \text{ where } s = 0, 1, 2, \ldots, s.$$

The general variation of the term to be summed is shown in the figure.

To estimate the value of P_s we write

$$r = x\sqrt{n},$$

and since the increment of r is unity,

$$r+1 = (x+\delta x)\sqrt{n},$$

so that $\delta x = 1/\sqrt{n}$. Thus

$$P_s = \sum_{x=-s/\sqrt{n}}^{x=s/\sqrt{n}} \frac{\delta x}{\sqrt{\pi}} e^{-x^2} = \frac{1}{\sqrt{\pi}} \int_{-s/\sqrt{n}}^{s/\sqrt{n}} e^{-x^2}\, dx, \text{ approximately,}$$

$$= \frac{2}{\sqrt{\pi}} \int_{0}^{s/\sqrt{n}} e^{-x^2}\, dx.$$

The function $\frac{2}{\sqrt{\pi}}\int\limits_0^x e^{-x^2}\,dx$ is known as the Probability Integral or the Error Function, and is denoted by Erf x (see Appendix).

Ex. 1. From a population containing equal numbers of boys and girls a sample of 1,800 is selected. Find the probability that the number of girls will differ from the number of boys by more than 100.

We seek first the probability that this excess will not occur. We have $2n = 1{,}800$, $s = 50$; then the probability that the number of girls is between $n+s = 950$ and $n-s = 850$ is

$$\frac{2}{\sqrt{\pi}}\int\limits_0^{50/\sqrt{900}} e^{-x^2}\,dx = \frac{2}{\sqrt{\pi}}\int\limits_0^{5/3} e^{-x^2}\,dx.$$

From the table (p. 197) we find that $\text{Erf}(5/3) = 0{\cdot}9816$.

Thus the probability that the difference is greater than this is

$$1-0{\cdot}9816 = 0{\cdot}0184, \quad \text{or} \quad 1{\cdot}8 \text{ chances in } 100.$$

Ex. 2. If we define a 'fair sample' of size $2n$ of a population of coins as one whose discrepancy from n heads is exceeded only in 5 cases out of 100, what is the discrepancy allowable in a fair sample?

Here we have to find s in terms of n from the definition that the probability of a fair sample is $5/100 = 0{\cdot}05$.

Thus $$\text{Erf}(s/\sqrt{n}) = \frac{2}{\sqrt{\pi}}\int\limits_0^{s/\sqrt{n}} e^{-x^2}\,dx = 0{\cdot}05.$$

From the table we find that $s/\sqrt{n} = 0{\cdot}044$.

Ex. 3. What should be the discrepancy such that as many cases have less than this as greater?

Here we require $$\text{Erf}(s/\sqrt{n}) = 0{\cdot}5,$$

whence $$s/\sqrt{n} = 0{\cdot}48.$$

Thus, if $2n = 800$, $s = 9{\cdot}6$, i.e. the range (390, 410) should include about half the number of cases.

Ex. 4. A penny is tossed 100 times, giving 45 heads and 55 tails. On the assumption that this is a sample of a large population containing equal numbers of heads and tails, find the percentage of cases in which a deviation at least as large as this will be expected.

We have $2n = 100$, $s = 5$, so that the probability of such cases is

$$1-\frac{2}{\sqrt{\pi}}\int\limits_0^{0{\cdot}707} e^{-x^2}\,dx = 1-0{\cdot}68262 = 0{\cdot}31738.$$

Hence the percentage of cases is about 32.

The General Case

We pass now to the general case in which the probability of a certain subclass of a given population is p. We have already shown (p. 62) that out of a sample of size n, the most probable number of members of the subclass is np or the least integer† greater than np if np is not itself an integer. We now seek the probability that in a large sample of size n the number r of members of the subclass differs by an amount x from the most probable number.

The probability of just r members occurring is

$$P = \frac{n!}{r!\,(n-r)!}p^r(1-p)^{n-r}.$$

Write $r = pn+x$, $n-r = (1-p)n-x$. Since n is large, r is large also, provided that x is small compared with np.

Using Stirling's formula and expanding in descending powers of n, we have

$$\begin{aligned}\log P &= \log n! + (pn+x)\log p + \{(1-p)n-x\}\log(1-p) - \\ &\qquad -\log(pn+x)! - \log\{(1-p)n-x\}! \\ &= -\tfrac{1}{2}\log 2\pi p(1-p)n - \frac{1}{2n}\left\{\frac{x^2}{p(1-p)} + \frac{x(1-2p)}{p(1-p)}\right\}\ldots.\end{aligned}$$

Thus $P = \dfrac{1}{\sqrt{\{2\pi(1-p)pn\}}}e^{-\frac{1}{2n}\{x^2+(1-2p)x\}/p(1-p)}$, approximately.

If $|x|$ is much greater than $|1-2p|$, we can neglect the term $(1-2p)x$ in comparison with x^2, in the exponent. We then obtain the approximation

$$P = \frac{1}{\sqrt{\{2\pi p(1-p)n\}}}e^{-x^2/2np(1-p)}$$

to the probability that a sample of large size n will contain $[pn]+x$ members of the subclass whose probability is p, where

$$pn \gg |x| \gg |1-2p|.$$

This result is also valid for $x = 0$, for which the probability is a maximum; that is, $[pn]$ is the most probable number of members of the subclass, and the probability that a sample of size n will have just this number is

$$\frac{1}{\sqrt{\{2\pi p(1-p)n\}}}.$$

Thus, the probability that a sample of size n will have a number of members of the subclass lying in the range $(pn-s, pn+s)$ is the sum of the probabilities that the sample will have

† This will be denoted by $[np]$.

precisely $[pn]+s$, $[pn]+s-1,\ldots$, $[pn]-s$, members of the subclass. Thus the required probability is

$$\sum_{x=-s}^{x=s} \frac{1}{\surd\{2\pi p(1-p)n\}} \exp\left\{-\frac{x^2+(1-2p)x}{2np(1-p)}\right\}.$$

We notice that

$$\sum_{-s}^{s} \exp\left\{-\frac{x^2+(1-2p)x}{2p(1-p)n}\right\}$$

$$= \sum_{-s}^{s} \exp\left\{\frac{-x^2}{2p(1-p)n}\right\}\left\{1-\frac{1-2p}{2p(1-p)}\frac{x}{n}+\ldots\right\}.$$

Since the summation extends to equal numbers of positive and negative terms, the second term in the brackets vanishes. Thus the probability required is approximately

$$\sum_{-s}^{s} \frac{1}{\surd\{2\pi p(1-p)n\}} \exp\left\{\frac{-x^2}{2np(1-p)}\right\}.$$

Write $y = x/\surd\{2np(1-p)\}$; then since x increases by unity, we have

$$y+\delta y = (x+1)/\surd\{2np(1-p)\},$$

so that

$$\frac{\delta y}{\surd\pi} = \frac{1}{\surd\{2\pi p(1-p)n\}}.$$

The summation then takes the form

$$\sum_{-s/\surd\{2np(1-p)\}}^{s/\surd\{2np(1-p)\}} e^{-y^2}\frac{\delta y}{\surd\pi} = \frac{2}{\surd\pi}\int\limits_{0}^{s/\surd\{2np(1-p)\}} e^{-y^2}\,dy, \text{ approximately.}$$

This result expresses the probability approximately in terms of the error function.

Thus

$$P = \operatorname{Erf}[s/\surd\{2np(1-p)\}].$$

Ex. 1. If there are 32 females to 30 males in the general population, what would be the most probable number, *ceteris paribus*, of women students in a university population of 1,800? What is the probability that the number of women students will be less than that number by 40?

The probability p of an individual being a female is $p = \frac{32}{62} = \frac{16}{31}$.

We have $n = 1{,}800$, $s = 40$, so that

$$s/\surd\{2np(1-p)\} = 40/30, \text{ approximately.}$$

From the table we find that Erf(4/3) = 0·94. Hence the probability that the women students exceed the men by less than 40 is very great.

Ex. 2. Defining a fair sample as one whose discrepancy s from np is in excess or in defect in only 10 per cent. of all cases, we have

$$\mathrm{Erf}[s/\sqrt{\{2np(1-p)\}}] = 0{\cdot}1.$$

From the table, $s = \sqrt{\{2np(1-p)\}} \times 0{\cdot}09.$

Thus; if $p = \frac{1}{10}$, $n = 5{,}000$, we obtain

$$s = \sqrt{(10^4 . \tfrac{1}{10} . \tfrac{9}{10})} \times 0{\cdot}09 = 2{\cdot}7.$$

Also $np = 500$. Thus a fair sample should, on this definition, have no more than 503 or no less than 497 of members belonging to the subclass.

EXAMPLES ON CHAPTER V

Ex. 1. If a penny is tossed 3 times, what is the probability of obtaining 2 heads?

Ex. 2. What is the probability of throwing an ace exactly once in 6 throws with a die?

Ex. 3. If m dice are thrown, show that the probability of obtaining an even number of aces is $\frac{1}{2}\{1+(\frac{2}{3})^m\}$.

Ex. 4. Drawings are made from a pack of 3 cards, of which 1 is red and 2 are black, and each time the card drawn is returned to the pack. If 10 such drawings are made, find the probability that n red cards will be chosen ($n = 0, 1, \ldots, 10$), and show that it is most probable that $n = 3$.

Ex. 5. Find the probability that in 8 throws of a die, the numbers 1, 3, 5 turn up 2, 3, 3 times respectively.

Ex. 6. A pack of $2n$ cards, n of which are red and n black, is divided into two equal parts, and a card drawn from each. Find the probability that the cards drawn are of the same colour, and compare with the probability that two cards drawn from the original pack should be of the same colour.

Ex. 7. A coin is tossed $m+n$ times ($m > n$). Show that the probability of at least m consecutive heads is $(n+2)/2^{m+1}$.

The required probability is the sum of the probabilities that there should appear exactly m, $m+1$, $m+2, \ldots, m+n$ consecutive heads. Now a series of m consecutive heads may begin at the first, second,..., $(n+1)$th throw; and since $m > n$, there cannot occur more than one such series. The probabilities of the first and last of these cases are evidently $1/2^{m+1}$, and of the others $1/2^{m+2}$. Thus the probability of a series of exactly m consecutive heads is

$$2/2^{m+1}+(n-1)/2^{m+2} = (n+3)/2^{m+2}.$$

Similarly, the probability of a series of $m+1$ consecutive heads is $(n+2)/2^{m+3}$, and so on, up to $m+n-2$. Finally, the probability of a series of exactly $m+n-1$ consecutive heads is $1/2^{m+n-1}$, and of $m+n$ consecutive heads is $1/2^{m+n}$.

Hence the required probability is

$$\frac{n+3}{2^{m+2}}+\frac{n+2}{2^{m+3}}+\ldots+\frac{5}{2^{m+n}}+\frac{1}{2^{m+n-1}}+\frac{1}{2^{m+n}}.$$

The first $n-1$ terms of this expression form an arithmetico-geometric series, the sum of which can be written down;[†] thus, we obtain for the probability the value $(n+2)/2^{m+1}$.

Ex. 8 (Pascal's problem). A and B play a game which must be either lost or won. If the probability that A wins any game is p, what is the probability that A wins m games before B wins n?

Evidently, the probability that B wins any game is $q = 1-p$. Now the required probability is that of A winning at least m games out of a series of $m+n-1$, that is, by Bernoulli's Theorem,

$$^{m+n-1}C_0 p^{m+n-1} + {}^{m+n-1}C_1 p^{m+n-2} q +$$
$$+ {}^{m+n-1}C_2 p^{m+n-3} q^2 + \dots + {}^{m+n-1}C_m p^m q^{n-1}.$$

Ex. 9. A bag contains m white and n black balls. If the balls are drawn out one by one, find the probability of drawing first a white and then a black, and so on, alternately, until all the balls remaining are of the same colour.

If m balls are drawn out at once, what is the probability that these are white?

Ex. 10. Four cards are drawn from a pack of 52; find the probability that they are all of different suits, (a) when each card is returned to the pack after the draw, (b) when it is not.

Ex. 11. Given n independent events $A_1, A_2, \dots, A_n$, whose respective probabilities are $p_1, p_2, \dots, p_n$, prove that the probability that at least one of the events happens is $\sum p_1 - \sum p_1 p_2 + \sum p_1 p_2 p_3 \dots$.

Ex. 12. With the notation of the previous example, show that the probability that the events $A_1, A_2, \dots, A_r$, and no more, happen is $p_1 p_2 \dots p_r (1-p_{r+1})(1-p_{r+2}) \dots (1-p_n)$. Hence find

(i) the probability that r (and no more) of the events happen;

(ii) the probability that r at least of the events happen.

Ex. 13. Out of a family of n offspring consisting of two equally probable types, r at least of one type are just as likely to occur as not: find the value of r.

The number r is determined by the equation

$$({}^nC_r + {}^nC_{r+1} + \dots + {}^nC_n)\frac{1}{2^n} = \frac{1}{2},$$

or
$$1+n+\frac{n(n-1)}{2!}+\dots+\frac{n(n-1)\dots(n-r+1)}{r!} = 2^{n-1}.$$

If n is even, there is no solution; but if n is odd, say $2m+1$, then $r = m+1$.

[†] See Chrystal, *Algebra*, ch. xx, 13.

CHAPTER VI

EXTENSION TO CONTINUOUS DISTRIBUTIONS

Definition

LET P_0P_1, $P_1P_2, \ldots$, $P_{l-1}P_l, \ldots$, $P_{n-1}P_n$ represent a series of n straight lines (or 'elements') to which the same measures of length δ have been attached. Suppose that they are joined end to end and that they are divided, for our purpose, into two classes: the first class L is to consist of those elements, l in number, lying to the left of P_l, and the second class R of those lying to the right. Then the probability that one of the set of elements shall be a member of L is l/n. We may arrive at this result in a different manner by inquiring what is the probability that a point selected anywhere in one of the elements, otherwise unspecified, shall lie in the class L; since such a point must lie in one of the elements, the required probability is l/n.

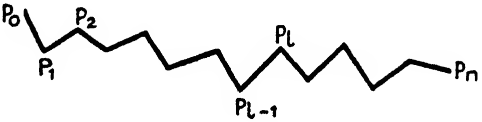


FIG. 4.

Now
$$\frac{l}{n} = \frac{l\delta}{n\delta} = \frac{\text{length of subclass } L}{\text{length of class } L+R}.$$

This is true no matter how many members the class and the subclass may contain, and however the successive elements are orientated with respect to one another.

Now let us suppose that to the total length P_0P_n a measure a has been attached and that to P_0P_l a measure b has been attached, so that $n\delta = a$ and $l\delta = b$; if δ is rational, then so are a and b. Let us proceed to the limit, making $n \to \infty$ and $\delta \to 0$.

It follows that, if $P_0P_lP_n$ is any continuous curve such that a, b are the measures adopted for the arc-lengths P_0P_n and P_0P_l, then the probability that a point known to lie on the arc P_0P_n shall lie on the arc P_0P_l is b/a. The probability that it shall lie on the arc P_lP_n is $1-(b/a)$.

If b and a are incommensurable (e.g. if $a = \surd 2$, $b = \surd 2$) it might appear that by no process of subdivision, in which each element has a rational measure, could an arc $P_0 P_l$ be obtained as the limit of a number of elementary straight lines. But, in fact, we may replace P by any rational point P of section in the element $P_{l-1} P_l$ since, on our definition of probability, it is immaterial where P_l lies in that element; and as the number n of subdivisions tends to infinity, the distance PP_l can be made to differ from zero by any assigned positive quantity. Thus the original proposition can be applied to irrational lengths of arc.

Analytically, if $y = f(x)$ is the equation to a curve passing through the points P_1, P_2 (having abscissae $x = a_1, a_2$) and if

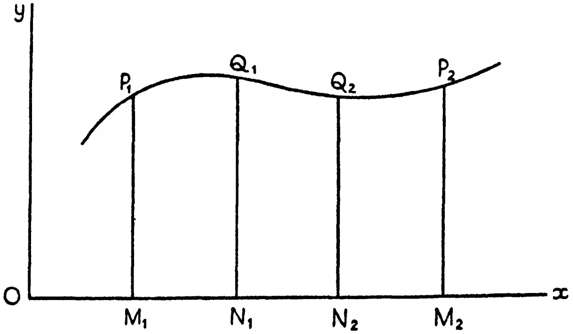


FIG. 5.

Q_1, Q_2 are internal points of the range (with abscissae $x = b_1, b_2$), then the probability that a point known to lie on the arc $P_1 P_2$ shall also lie on the arc $Q_1 Q_2$ is

$$\frac{\int_{Q_1}^{Q_2} ds}{\int_{P_1}^{P_2} ds} = \frac{\int_{b_1}^{b_2} \sqrt{\{1+f'(x)^2\}}\, dx}{\int_{a_1}^{a_2} \sqrt{\{1+f'(x)^2\}}\, dx},$$

where s is the arc-length of the curve measured from some fixed point.

If $M_1 N_1 N_2 M_2$ are the feet of the ordinates at the four points, as shown, then the probability that a point known to lie in the range $M_1 M_2$ shall also lie in the range $N_1 N_2$ is $N_1 N_2 / M_1 M_2$.

Ex. 1. As an illustration of the above results, consider a semicircle of radius r, bounded by a diameter $M_1 M_2$. First let us find the expectation of the height of the ordinate PN drawn from a point P known to lie on the arc M_1M_2 but otherwise unspecified. If C is the centre of the semicircle and the angle PCN is θ, then the probability that P lies on an elemental arc of measure $r\,d\theta$ is $r\,d\theta/\pi r$, by definition. And since $PN = r\sin\theta$, the expected height of the ordinate is

$$\int\limits_0^\pi r\sin\theta\frac{r\,d\theta}{\pi r} = \frac{r}{\pi}\int\limits_0^\pi \sin\theta\,d\theta = \frac{2r}{\pi}.$$

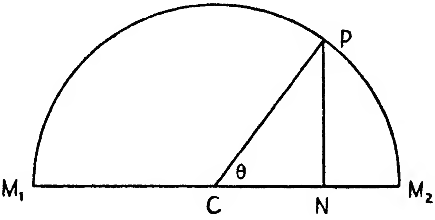


Fig. 6.

Now let us find the expected height of the ordinate PN erected at a point N known to lie in M_1M_2 but otherwise unspecified. If $CN = x$, then $PN = \sqrt{(r^2-x^2)}$, and the expected height of the ordinate is

$$\int\limits_{-r}^{r} \sqrt{(r^2-x^2)}\frac{dx}{2r} = \frac{1}{r}\int\limits_0^r \sqrt{(r^2-x^2)}\,dx = \tfrac{1}{4}\pi r.$$

Note the difference between these two results: to what is it due?

Ex. 2. A line PQ is bisected at R. Two points S, T are known to lie on PQ. Find the probability that (1) they are on opposite sides of R, (2) they are on the same side of R, (3) they are both to the right of R.

Applications to Weighted Probabilities

Questions of geometrical probability arise in which, as in the example previously considered (p. 61), some bias has to be allowed for; thus, in the above formulation of our definition, let us suppose quite generally that the element $\overline{P_0P_1}$ is 'weighted' with a number p_1, that the element $\overline{P_1P_2}$ is weighted with p_2,..., and that $\overline{P_{l-1}P_l}$ is weighted with p_l. Then the probability that an element of the class shall belong to L is now

$$\sum_{l=1}^{l} p_l\overline{P_{l-1}P_l}\Big/\sum_{l=1}^{n} p_l\overline{P_{l-1}P_l}.$$

Similarly, in the case of the continuous curve P_0P_n, if a point P, whose position on the arc P_0P_n is defined by a measure s of arc-length, is weighted by an amount $p(s)$, where $p(s)$ is some function of s, then the probability that P shall lie on the arc P_0P_r is

$$\int_{P_0}^{P_r} p(s)\,ds \Big/ \int_{P_0}^{P_n} p(s)\,ds.$$

Extension to Two Dimensions

Suppose that a plane is divided into rectangles by lines drawn parallel to the coordinate axes Ox, Oy. Consider a polygon $ABCDEF$ bounded by sides of these rectangles, to each of which a measure α of area has been attached, and suppose that it contains a of these rectangles. Let $PQRST$ be a polygon lying within $ABCDEF$ and bounded by sides of the same rectangles, of which it contains b, suppose. Then if a rectangle is known to be one of the class a, the probability that it shall also be one of the subclass b is

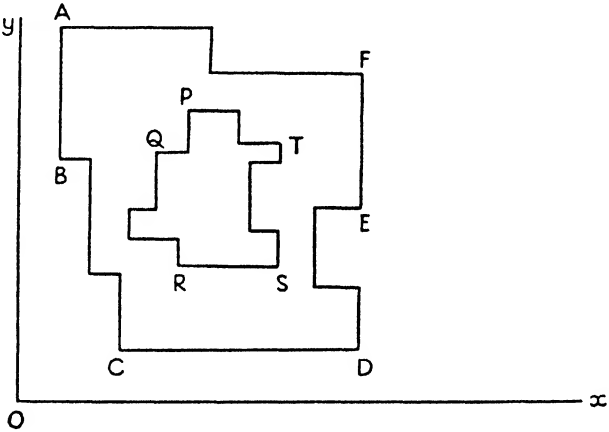


Fig. 7.

$$\frac{b}{a} = \frac{b\alpha}{a\alpha} = \frac{\text{area of polygon } PQRST}{\text{area of polygon } ABCDEF}.$$

We may now pass, by a discussion analogous to the preceding, to the following

Theorem. *If S is a simple closed curve, of area a, containing a simple closed curve S' of area b, then the probability that a point lying in the region enclosed by S shall also lie in the region enclosed by S' is b/a.*

For the procedure by which this result is established we may refer the reader to the usual method of obtaining the formula for the area enclosed by a curve. By subdividing the area S into a meshwork of elementary rectangles we thus obtain for the required probability the formula

$$\iint_{S'} dxdy \Big/ \iint_{S} dxdy,$$

in which the integrals are taken up to the boundaries S and S'.

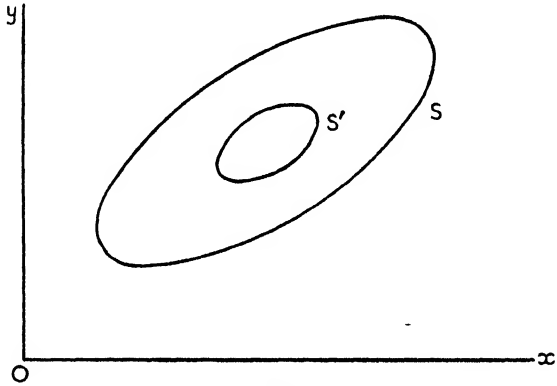


FIG. 8.

If the problem is one of weighted probability, we suppose that to a point with coordinates (x, y) situated within S, the weight attached is some function $f(x, y)$ of its coordinates. Then the probability required is

$$\iint_{S'} f(x, y)\, dxdy \Big/ \iint_{S} f(x, y)\, dxdy.$$

Discrete and Continuous Entities

To illustrate the passage from a problem in probability dealing with discrete entities to one concerning a continuous medium, consider the following:

A population consists of elements forming two subclasses b and w in the proportion of $1 : T-1$. The number of elements in any sample of magnitude T is n. From this population is drawn a sample of total magnitude t; we require to find the probability that in this sample there is *no* member of the subclass b.

If we assume that nt/T is an integer, it follows that the number of elements in the sample of magnitude t is nt/T. And since the proportion of b to the whole population is $1:T$, the probability that an element in a sample of magnitude T belongs to b is $1/n$; thus the probability that it does *not* belong to b is $1-1/n$. If we now consider the sample t, containing nt/T elements, the probability that none of them belongs to b is

$$\left(1-\frac{1}{n}\right)\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{1}{n}\right)\ldots \text{ to } \frac{nt}{T} \text{ factors} = \left(1-\frac{1}{n}\right)^{nt/T},$$

$$= \left(1-\frac{1}{n}\right)^{-n(-t/T)}$$

If n is sufficiently large,† $\left(1-\frac{1}{n}\right)^{-n}$ is approximately e, where

$$e = 1+\frac{1}{1!}+\frac{1}{2!}+\ldots = 2{\cdot}71828\ldots.$$

Hence the required probability is approximately $e^{-t/T}$.

If the original population be considered as a continuous one, e.g. a volume of water or an interval of time or space, then the number n of elements in the sample may be made arbitrarily large, and whatever the value of t/T, provided it is rational, we can always assume that nt/T is arbitrarily large and integral. Thus, we can assert the following:

THEOREM. *If in any continuously varying process (varying e.g. with respect to time, space, or volume) a certain characteristic is present to the extent of one in T units, then the probability that the characteristic does not occur in a sample of t units is $e^{-t/T}$.*

Ex. 1. It is known that 100 litres of water have been polluted with 10^6 bacteria. If 1 c.c. of water is drawn off, what is the probability that the sample is not polluted?

Since 100 litres $= 10^5$ c.c., it follows that $T = \frac{10^5}{10^6} = 10^{-1}$. Also $t = 1$; so that the required probability is

$$e^{-10} = 0{\cdot}000045, \text{ approximately.}$$

Ex. 2. An aircraft company carries on the average P passengers M miles for every passenger killed. What is the probability of a passenger completing a journey of m miles in safety?

The fatal accidents occur once in PM passenger miles. Hence the probability that an accident should not occur in m given passenger miles is approximately $e^{-m/MP}$.

† For example, if $n = 1000$, the error in replacing $\left(1-\frac{1}{n}\right)^{-n}$ by e does not affect the second decimal place.

Ex. 3. Estimate from this result an apparently reasonable premium to pay in order that, should a passenger be killed on such a flight, his heir should receive £10,000.

Ex. 4. If during the week-end road traffic 100 cars per hour pass along a certain road, each taking 1 minute to cover it, find the probability that at any given instant no car will be on this road.

Evidently no car must have entered the road during the previous minute. But on the average a car enters every $\frac{3600}{100} = 36$ sec.

Thus the required probability is $e^{-60/36} = e^{-5/3}$.

Ex. 5. In a completed book of 540 pages 624 typographical errors occur. What is the probability that 4 specimen pages selected for advertisement are free from errors?

Ex. 6. A series of cars of the same length and with the same speed proceed along a certain road, one every T seconds; and another series of cars identical in length and speed with the first, proceed along a road meeting the first road at right angles, one car passing every T' seconds. If a car takes t seconds to pass an observer, find the probability that there should be no collisions in an interval of time t.

By 'collision' we mean in this case the situation of some portions of two cars, at the cross-roads, at the same instant.

The required probability is evidently the sum of the following separate probabilities:

(1) the probability that no car on the first road is passing the cross-roads in the interval t, and that a car on the second road is passing the cross-roads in that interval;

(2) the probability that no car on the second road is passing in the interval and that a car on the first road is passing;

(3) the probability that no car passes the cross-roads on either road during the interval.

Hence the probability is

$$e^{-t/T}(1-e^{-t/T'})+e^{-t/T'}(1-e^{-t/T})+e^{-t/T}e^{-t/T'} = e^{-t/T}+e^{-t/T'}-e^{t(1/T+1/T')}.$$

Ex. 7. Criticize the following statements:

(1) The sun rises once per day; hence the probability that it will not rise to-morrow is e^{-1}.

(2) The probability that it will rise *at least once* is $1-e^{-1}$.

The 'Random Walk' Problem

We begin with the simple case in one dimension. An individual is constrained to move backwards and forwards in a straight line, each step being of length l, it being at each stage equally probable that the step will be taken forward or backward. We inquire what is the probability that after n steps his displacement will lie between a and $a+da$, where n is large.

Let $a = ml$; then clearly we have to calculate the probability P that out of n steps $\frac{1}{2}(n+m)$ will be forward and $\frac{1}{2}(n-m)$

backward. The probability of each step being $\frac{1}{2}$, by Bernoulli's Theorem

$$P = \frac{n!}{[\frac{1}{2}(n+m)]!\,[\frac{1}{2}(n-m)]!}\,\frac{1}{2n},$$
$$= \left(\frac{2}{\pi n}\right)^{\frac{1}{2}} e^{-m^2/2n},$$

by applying Stirling's approximation for large values of n.

Thus the probability that the displacement will lie between a and $a+da$ is

$$\frac{1}{\sqrt{(2\pi n l^2)}}\, e^{-a^2/2nl^2}\, da.$$

The mean square distance σ^2 is then given by†

$$\sigma^2 = \frac{1}{\sqrt{(2\pi n l^2)}} \int\limits_{-\infty}^{+\infty} a^2 e^{-a^2/2nl^2}\, da = nl^2$$

or $\sigma = l\sqrt{n}$, and the required probability is

$$\frac{1}{\sigma\sqrt{(2\pi)}}\, e^{-a^2/2\sigma^2}\, da.$$

We pass now to the two-dimensional case.

A man walks a distance $OO_1 = l_1$ from a point O in any direction and then walks a distance $O_1\,O_2 = l_2$ in any direction; required the probability that the final point O_2 falls within distances r_1 and r_2 of O, where $r_2 > r_1$.

Draw a circle of radius l_2 about O_1 cutting the circles of radii r_1 and r_2 about O at P and Q. O_2 may fall anywhere on the circle with centre O_1, and it will satisfy the required conditions if it falls on the arc PQ. Hence the required probability is

$$p = \frac{PQ}{\pi l_2} = \frac{\angle\, PO_1\,Q}{\pi}$$
$$= \frac{1}{\pi}\left[\cos^{-1}\frac{l_1^2+l_2^2-r_2^2}{2l_1\,l_2} - \cos^{-1}\frac{l_1^2+l_2^2-r_1^2}{2l_1\,l_2}\right].$$

Ex. 1. If $l_1 = l_2 = l$, the probability that the final position lies between a distance r and $r+dr$ from O is

$$\frac{2}{\pi}\,\frac{dr}{(4l^2-r^2)}.$$

Ex. 2. Two points P and Q are at distance l_1 apart. A man walks from Q in a straight line to a point R which is then found to be a distance

† See Chap. VIII.

l_2 from P. What is the probability that the distance QR lies between l_3 and l_4?

What is the probability that QR lies between λ and $\lambda+d\lambda$?

ILLUSTRATIVE EXAMPLES

Ex. 1. Two points are selected in a line AC of length a, so as to lie on opposite sides of its mid-point O.

Find the probability that the distance between them is less than $\frac{1}{3}a$.

Let P and Q be the points and let $OP = x$, $QO = y$.

We thus require

$$x+y < \tfrac{1}{3}a.$$

The conditions of the problem require further that $x < \frac{1}{2}a$, $y < \frac{1}{2}a$.

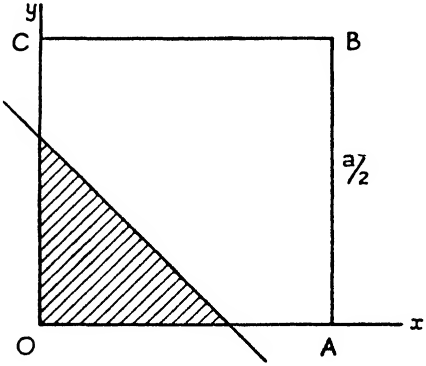


Fig. 9.

If we represent x and y by Cartesian coordinates, it is clear that x and y may lie anywhere within the square shown, while the values of x and y which satisfy the condition $x+y < \frac{1}{3}a$ lie in the shaded area.

Hence the required probability $= \dfrac{a^2}{18}\Big/\dfrac{a^2}{4} = \dfrac{2}{9}$.

Ex. 2. A line of given length is divided into three parts. Find the probability that these will form the sides of a triangle.

Let AB be the line, of length a, and let the three parts be x, y, and $a-(x+y)$.

Then we require

$$x+y > a-(x+y),$$
$$x+a-(x+y) > y,$$
$$y+a-(x+y) > x.$$

These conditions are equivalent to $x+y > \frac{a}{2}$, $x < \frac{a}{2}$, $y < \frac{a}{2}$. In any case we have the condition $x+y < a$.

Hence, if we represent x and y by Cartesian coordinates, as before and the lines BD, AE by $x+y = a$, $x+y = a/2$, respectively, the required probability is evidently

$$\frac{\text{area } ACE}{\text{area } OBD} = \frac{1}{4}.$$

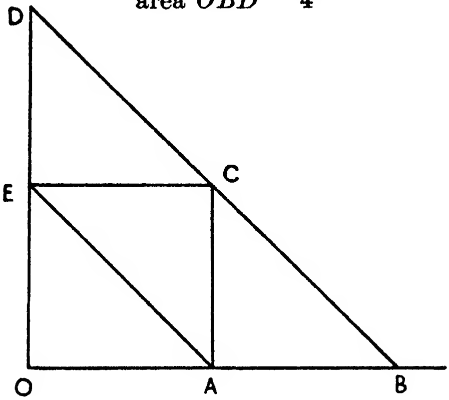


Fig. 10

Ex. 3. Find the probability that the roots of the equation $x^2+2px+q = 0$, where $-P \leqslant p \leqslant P$ and $-Q \leqslant q \leqslant Q$, should be real.

Let p and q be represented by Cartesian coordinates, so that they are restricted to lie in the rectangle shown. The condition

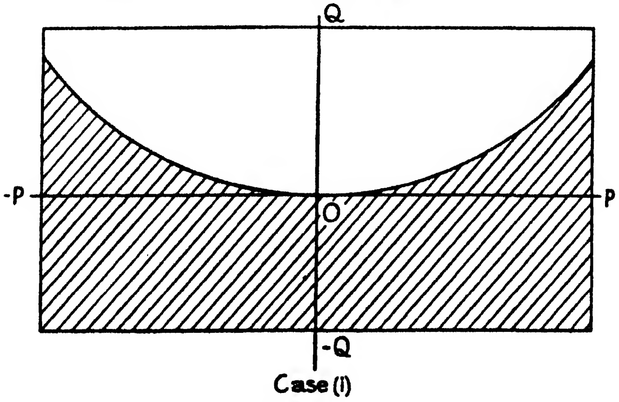
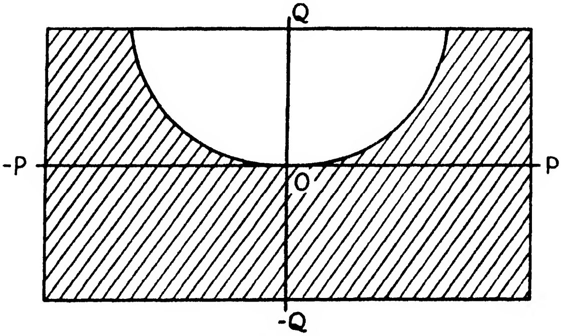


Fig. 11.

for the reality of the roots is $p^2 \geqslant q$; thus p and q must be such that the point (p, q) lies on the lower side of the parabola $y = x^2$. There are two cases to distinguish, according as $P^2 \leqslant Q$ or $P^2 > Q$. (i) If $P^2 \leqslant Q$, the shaded area $= 2\int\limits_0^P y\,dx + 2PQ$, the integral being taken along the parabola.



Case (ii)

FIG. 12.

Thus the area is $\frac{2P^3}{3} + 2PQ$, and the required probability is therefore

$$\left(\frac{2P^3}{3} + 2PQ\right)\Big/4PQ = \frac{1}{2} + \frac{P^2}{6Q}.$$

(ii) If $P^2 > Q$, the shaded area

$$= 4PQ - 2\int\limits_0^Q x\,dy$$

$$= 4PQ - \frac{4Q^{\frac{3}{2}}}{3},$$

and the probability is

$$(4PQ - \tfrac{4}{3}Q^{\frac{3}{2}})/4PQ = 1 - \frac{Q^{\frac{1}{2}}}{3P}.$$

Ex. 4. If two points P, Q are taken in a circle, what is the probability that the circle with centre P and radius PQ will lie inside the original circle?

Let the radius of the original circle be a (Fig. 13); then the probability that P lies in an annulus of breadth dx at a distance x from the centre O is

$$\frac{2\pi x\,dx}{\pi a^2} = \frac{2x\,dx}{a^2}.$$

The second circle will lie inside the first if $PQ < PN$, where

$PN = a-x$. Thus Q may lie anywhere within a circle with centre P and radius $a-x$. Hence the required probability is

$$\int_0^a \frac{\pi(a-x)^2}{\pi a^2}\,\frac{2x\,dx}{a^2} = \frac{2}{a^4}\int_0^a (a^2x-2ax^2+x^3)\,dx$$

$$= \frac{2}{a^4}\left[\frac{a^4}{2}-\frac{2a^4}{3}+\frac{a^4}{4}\right] = \frac{1}{6}.$$

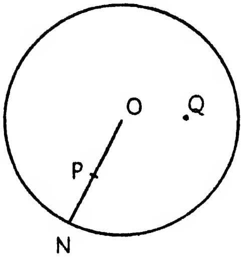


FIG. 13.

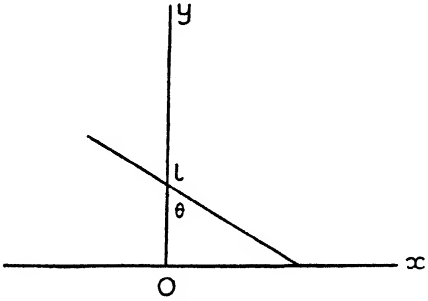


FIG. 14.

Ex. 5. *Buffon's problem.* A smooth table is ruled with parallel lines at distance a apart. A needle of length $l < a$ is dropped on the table. What is the probability that it will cross one of the lines?

Take one of the parallel lines for x-axis and any perpendicular to it for y-axis (Fig. 14). The probability that the centre of the needle has an ordinate lying between the limits y and $y+dy$ is dy/a; and the probability that the inclination of the needle to Oy should be between θ and $\theta+d\theta$ is $\dfrac{d\theta}{\pi}$. Hence the probability that the needle will cross Ox is

$$\iint \frac{dy d\theta}{a\pi},$$

where the double integral is taken over the range of values of y and θ for which the needle will cross Ox. The possible values of y are evidently given by $|y| \leqslant \frac{1}{2}l\cos\theta$, and θ lies in the range $-\frac{1}{2}\pi \leqslant \theta \leqslant \frac{1}{2}\pi$. Thus, from Fig. 15, where DEA is the curve $y = \frac{1}{2}l\cos\theta$ and AB is of length $\frac{1}{2}a$, the required

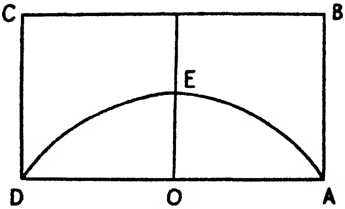


FIG. 15.

probability is
$$\frac{\text{area } AED}{\text{area } ABCD} = \frac{l}{\frac{1}{2}a\pi} = \frac{2l}{a\pi}.$$

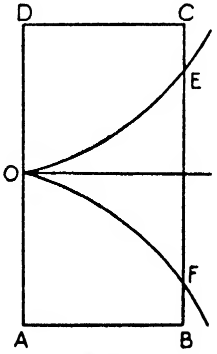
Ex. 6. Consider the same problem in the case where $l > a$.

Ex. 7. A point P is chosen on a line AB of length $2a$. What is the probability that $AP.PB$ should exceed λa^2, where λ is a given positive number ?

Ex. 8. A point is chosen on each of two adjacent sides of a square. Show that the average area of the triangle formed by the sides of the square and the line joining the two points is one-eighth of the area of the square.

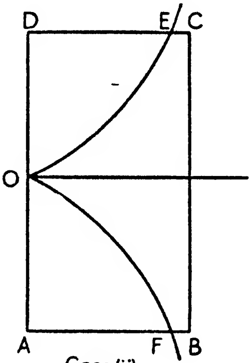
Ex. 9. Three points are chosen on the circumference of a circle. What is the probability that they lie on the same semicircle ?

Ex. 10*. Find the probability that the equation
$$x^{2n+1} = (2n+1)px+2nq,$$
where n is a positive integer and $0 \leqslant p \leqslant P$, $-Q \leqslant q \leqslant Q$, should have three of its roots real.



Case (i)

FIG. 16.



Case (ii)

FIG. 17.

By plotting the curves $y = x^{2n+1}$, $y = (2n+1)px+2nq$, it is easily seen that the condition for reality of the roots is $p^{2n+1} \geqslant q^{2n}$.

We now represent p and q by Cartesian coordinates (x, y), whence, as in Ex. 3, it follows from the diagrams shown that the required probability is $\dfrac{\text{area } OEF}{\text{area } ABCD}$. Thus two cases arise. In Case (i), the area
$$OEF = 2\int_0^P x^{\frac{2n+1}{2n}}\,dx = \frac{4n}{4n+1}P^{\frac{4n+1}{2n}},$$

so that the probability is

$$\frac{2n}{4n+1}\,\frac{P^{\frac{2n+1}{2n}}}{Q}.$$

In Case (ii), the area $OEF = 2PQ-2\int\limits_0^Q y^{\frac{2n}{2n+1}}\,dy$

$$= 2PQ-\frac{2(2n+1)}{4n+1}\,Q^{\frac{4n+1}{2n+1}},$$

and the probability is therefore $1-\dfrac{2n+1}{4n+1}\,\dfrac{Q^{\frac{2n}{2n+1}}}{P}$.

Ex. 11. Find the probability that the solutions of the simultaneous differential equations

$$\frac{dx}{dt}+bx+\frac{dy}{dt} = 0,$$

$$2(a-b)x+\frac{dy}{dt}+by = 0,$$

where $0 < a < A$, $0 < b < B$, represent decaying oscillations.

Eliminating y from the equations, we obtain for x the equation

$$\frac{d^2x}{dt^2}+2(2b-a)\frac{dx}{dt}+b^2x = 0.$$

For a decaying oscillation we requiie $a > 2b$ and $(2b-a)^2 < b^2$. This latter condition is equivalent to $(3b-a)(b-a) < 0$, so that either

(i) $3b < a,\ b > a$, or (ii) $3b > a,\ b < a$.

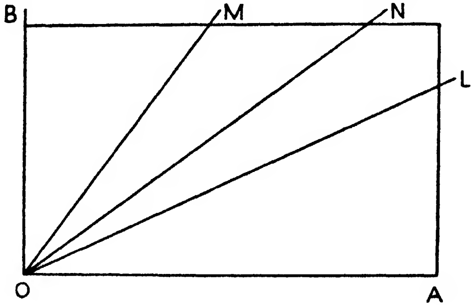


FIG. 18

If we represent a and b by Cartesian coordinates (x, y), their total field of variation is the rectangle bounded by the axes and $x = A$, $y = B$. For the conditions of the problem a and b

must be represented by values of (x,y) which lie between $y = 0$ and $y = \frac{1}{2}x$ (the line ON), and also between $y = \frac{1}{3}x$ and $y = x$ (the lines OL, OM); it is clear that the condition (i) cannot be fulfilled. If, for example, we suppose that $B > 2A$, the required probability is evidently $\frac{1}{2}(\frac{1}{2}A^2 - \frac{1}{3}A^2)/AB = \dfrac{A}{12B}$.

Ex. 12. What is the probability that the second figure in a table of square roots of x is n, if x ranges from 0 to 1 and is tabulated at equal intervals?

Let the first figure for any x be m; then for success we require

$$0{\cdot}mn \leqslant \sqrt{x} < 0{\cdot}m(n+1)$$

or
$$m+\frac{n}{10} \leqslant 10\sqrt{x} < m+\frac{n+1}{10},$$

where m may be 0, 1,..., 9.

Thus out of the total range of x within which it falls, viz., 0–1, the second figure in $\sqrt{x}$ will be n if x falls in any one of the intervals

$$\frac{1}{100}\left[\left(m+\frac{n+1}{10}\right)^2-\left(m+\frac{n}{10}\right)^2\right] = \frac{1}{10{,}000}(20m+2n+1),$$

corresponding to $m = 0, 1,..., 9$.

Hence the required probability is

$$P = \frac{1}{10{,}000}\sum_{m=0}^{m=9}(20m+2n+1).$$

Thus for $n = 0$, $P = 0{\cdot}0091$ and for $n = 9$, $P = 0{\cdot}109$.

Ex. 13. What is the probability that when $\log_e x$ is tabulated for $x = 1$ to $x = 0$, at equal intervals of x, the second figure in the table will be 2?

EXAMPLES ON CHAPTER VI

Ex. 1. A defective measuring instrument slips one scale division each time it is used. Find the probability that after being used 100 times it will be no more than 6 divisions from the zero reading.

Ex. 2. Trains leave a station at 3, 5, 8, 10, 13,... minutes past the hour. Find the probability that a passenger arriving at the station has to wait less than a minute for a train.

Ex. 3. A point P is chosen on a line AB. What is the probability that $AP : PB > \lambda$?

Ex. 4. Two points are taken in a circle. Find the probability that the perpendicular from the centre on the line joining them does not pass between them.

Ex. 5. On a chess-board, the squares of which are of side a, there is thrown a coin of diameter b, so as to lie entirely on the board, which includes a border of width c. Find the probability that it will lie entirely on one square ($a > b > c$).

Ex. 6. A floor is paved with tiles, each tile being a parallelogram such that the distances between pairs of opposite sides are a and b respectively, the length of the diagonal being l. A stick of length c falls on the floor parallel to this diagonal. Show that the probability that it will lie entirely on one tile is $\left(1-\frac{c}{l}\right)^2$.

If a circle of diameter d is thrown on the floor, show that the probability that it will lie on one tile is $\left(1-\frac{d}{a}\right)\left(1-\frac{d}{b}\right)$.

Ex. 7. A sheet of perforated zinc in the form of a square 22 cm. in width is covered with ten rows and ten columns of holes each 1 cm. in diameter, the centres in the rows and the columns being evenly spaced at intervals of 2 cm.

What is the probability that a particle of sand (considered as a point) blown against the zinc sheet will pass through to the other side?

What is the probability that a small shot of diameter $\frac{1}{2}$ cm. fired against the zinc without sufficient force to penetrate the metal will pass through one of the holes?

Ex. 8. A disk of wood of radius R and thickness d is cut so that it finally consists of four blades or sectors, each of 30°, radiating from the centre and evenly spaced. The disk is then set spinning with angular velocity ω about an axis through the centre at right angles to the disk; a shot is fired with velocity V parallel to and at distance $r < R$ from the axis. Find the probability that the shot will pass without damaging the blades of the disk.

Ex. 9. A point P lies inside a circle of diameter AB. What is the probability

(1) that $\pi > \angle APB > \alpha > \frac{1}{2}\pi$,

(2) that $\frac{1}{2}\pi > \angle PAB > \alpha > 0$,

where α is a given angle?

Ex. 10. Three chords are drawn through the same point of a circle. What is the probability that all three lines cut the same semicircle?

Ex. 11. A particle oscillates harmonically with period T between two points A and B distant $2a$ apart. What is the probability that during a small interval of time t the particle will be found within a small distance b of the point B?

Ex. 12. A raindrop falls steadily down a window-pane of total height H. At every distance h a grease spot deflects it by an amount d to the right or left. What is the probability that by the time it reaches the bottom it will have been deflected from its original direction of descent by an amount D?

CHAPTER VII

THE THEORY OF ARRANGEMENTS (2)

In the following theorems we are dealing with a series of problems that can perhaps be described best in this way. Let there be a row of pigeon-holes into which it is proposed to place a set of objects which may or may not differ from one another. The result of the distribution may be that some pigeon-holes contain objects and some do not. Thus the number of ways in which a distribution can be effected will depend upon two factors:

(1) Whether the order of the pigeon-holes, even including blanks, is taken into account.

(2) Whether the order of the objects within the pigeon-holes is taken into account.

The set of objects in a pigeon-hole will be called a *group* or a *parcel* according as the order of the objects is or is not taken into consideration. Unless otherwise stated, it is to be assumed throughout that the order of the pigeon-holes is significant.

Suppose that we are given n different objects in a row and that these are divided by $r-1$ partitions into groups which may range in size from 0 to n. In how many ways can this division be accomplished? Altogether, counting objects and partitions, we have $n+r-1$ entities, and if these are permuted among themselves we shall obtain the required number N of distributions, provided we make allowance for the fact that the interchange of two partitions does not alter the result. Thus we have to permute $n+r-1$ objects among themselves, $r-1$ of them being alike; so that, by the theorem (p. 42),

$$N = (n+r-1)!/(r-1)! = r(r+1)...(r+n-1).$$

Hence

Theorem I. *The number of ways in which n different objects can be arranged in r or fewer groups is* $r(r+1)...(r+n-1)$.

Ex. 1. Show that there are 6 ways of displaying 3 flags on 2 masts, when all the flags must be displayed but both masts need not be used.

Ex. 2. Show by means of Stirling's theorem that when n is large compared with r, the value of N in Theorem I is $\sqrt{(2\pi)}n^{n+r-\frac{1}{2}}e^{-n}/(r-1)!$, approximately.

Now let us impose the restriction that each of the r groups must contain at least *one* object. To find the number of ways in which the distribution can be made, we begin by selecting r of the n objects and placing one in each of the r compartments; since the objects are all different, this selection can be made in ${}^nP_r$ ways. For each such arrangement the problem now resolves itself into the preceding, for there remain $n-r$ objects to be distributed into r or less groups. Hence the total number N of ways is given by

$$N = {}^nP_r . r(r+1)...(r+n-r-1) = n!(n-1)!/(n-r)!(r-1)!.$$

Thus,

THEOREM II. *The number of ways in which n different objects can be arranged in exactly r groups is $n!(n-1)!/(n-r)!(r-1)!$.*

We note that when $n = r$, this reduces to $n!$, as expected.

Ex. 1. A builder has been asked to deliver 10 different consignments of materials on 4 successive days, at certain specified times. If he omits to record the details of the order in which the materials should be sent, what is the probability that he executes the order correctly?

Ex. 2. Applying Stirling's theorem to the result of Theorem II when n is large compared with r, show that the approximate value of N is

$$\sqrt{(2\pi)}n^{n+r-\frac{1}{2}}e^{-n}/(r-1)!$$

as in Theorem I, Ex. 2.

Ex. 3. By estimating the approximations in Theorems I and II to a higher degree of accuracy, determine the proportion of the total number of ways which arise from the assumption that fewer than r groups may be employed.

The last proposition can easily be generalized. If we wish to arrange n different objects into r groups so that each group contains at least s objects, we begin by selecting rs objects and placing s in each of the r groups. Since this selection can be made in ${}^nP_{rs}$ ways, we have the result:

THEOREM III. *The number of ways in which n different objects can be arranged in r groups, each of which contains at least s objects, is ${}^nP_{rs}r(r+1)...(r+n-rs-1)$.*

Now suppose that the n objects which we wish to arrange in r different groups are identical. This means, of course, that we are now dealing with parcels instead of groups. We begin by placing the objects in a row—there will then be $n-1$ gaps between them. If we indicate $r-1$ of these gaps we shall have separated the objects into parcels, each parcel containing at least one object. Thus the number of ways of forming such parcels is the number of ways of indicating $r-1$ gaps among the $n-1$, i.e. ${}^{n-1}C_{r-1}$. Hence,

THEOREM IV. *The number of ways in which n identical objects can be arranged in r different parcels is* $N = (n-1)!/(r-1)!(n-r)!$ Note that, by the method† of Theorem XII, this number can be obtained as the coefficient of x^{n-r} in

$$(x^0+x^1+\ldots+x^{n-r})^r = (1-x^{n-r+1})^r/(1-x)^r,$$

i.e. in $(1-x)^{-r}$. Thus the coefficient is ${}^{n-1}C_{r-1}$, as before.

Ex. 1. During a period of shortage, n tons of coal have to be distributed among r factories. What is the probability that a specified factory is supplied with exactly m tons?

By Theorem IV, the total number of ways in which the n tons can be supplied is $(n-1)!/(r-1)!(n-r)!$.

If m tons are given to the specified factory, we have $n-m$ tons left to distribute among the remaining $r-1$ factories, and this distribution can be effected in $(n-m-1)!/(r-2)!(n-m-r+1)!$ ways. Hence the required probability is $\dfrac{(r-1)(n-r)(n-r-1)\ldots(n-r-m+2)}{(n-1)(n-2)\ldots(n-m)}$. Thus, if $n = 10$, $r = 4$, $m = 3$, the probability is $5/28$.

Ex. 2. If n is large compared with r, show that the number of arrangements obtained in Theorem IV is approximately $n^{r-1}/(r-1)!$.

Ex. 3. Prove that the value of r for which N is greatest is the smallest integer not less than $\frac{1}{2}n$.

From the last theorem we can find the number of arrangements into r or less parcels. For the number of such arrangements is the number of ways in which $n+r-1$ objects can be distributed into r parcels, each containing at least one, whence

THEOREM V. *The number of ways in which n identical objects can be arranged in r or fewer parcels is* $(n+r-1)!/(r-1)!\,n!$.

COROLLARY. *The number of ways in which n identical objects can be arranged in r parcels, none of which contains less than q objects, is* ${}^{n-rq+r-1}C_{r-1}$.

For we place q objects in each of the r parcels, leaving $n-rq$ objects to be arranged in r or less parcels.

Ex. 1. n nuts are thrown daily into a cage containing r squirrels. If a squirrel to survive must have a ration of m nuts at least per day, and if in the struggle some get more than their share and others less, find the probability that a certain squirrel will survive.

† See p. 98.

Ex. 2. Suppose that $n = 5$, $r = 2$, $q = 1$, and let the five objects be denoted by letters a. Then the number of arrangements is evidently a, a^4; a^2, a^3; a^3, a^2; a^4, a, i.e. four.

Given n different objects we inquire in how many ways they can be distributed into r or less groups not necessarily using all the n objects.

Suppose we select x of the objects—such a selection can be made in ${}^nC_x$ ways—and then distribute these objects among themselves, as in Theorem I. In this way we obtain ${}^nC_x r(r+1)...(r+x-1)$ distributions; and since x may vary from 0 to n, the required number of ways is

$$N = \sum_{x=0}^{n} {}^nC_x r(r+1)...(r+x-1) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} \frac{(r+x-1)!}{(r-1)!}$$

$$= \frac{n!}{(r-1)!} \sum_{x=0}^{n} \frac{(r+x-1)!}{x!(n-x)!}.$$

Now let us form the product of the two series

$$e^x = 1+x+\frac{x^2}{2!}+...+\frac{x^n}{n!}+...,$$

$$(1-x)^{-r} = 1+rx+\frac{r(r+1)}{2!}x^2+..., \quad \text{where } x < 1.$$

The coefficient of x^n in this product is

$$\frac{1}{n!}+\frac{1}{(n-1)!}\frac{r}{1!}+\frac{1}{(n-2)!}\frac{r(r+1)}{2!}+...+\frac{r(r+1)...(r+n-1)}{n!}$$

$$= \frac{1}{(r-1)!}\left[\frac{(r-1)!}{n!}+\frac{r!}{1!(n-1)!}+\frac{(r+1)!}{2!(n-2)!}+...\right].$$

On comparing this expression with the above value of N we obtain the theorem:

Theorem VI. *The number of ways in which n different objects can be distributed into r or fewer groups, not necessarily using all the n objects, is the coefficient of x^n in the expansion of $n!\,e^x(1-x)^{-r}$.*

Ex. Thus, if $n = 2$, $r = 2$, the number of arrangements is the coefficient of x^2 in $2e^x(1-x)^{-2}$, i.e. in

$$2(1+x+\tfrac{1}{2}x^2+...)(1+2x+3x^2+...).$$

Hence the required number is 11. As a verification we find that the number of arrangements of two objects a and b is given by the scheme:

$$a,0;\quad b,0;\quad ab,\ 0;\quad ba,\ 0;\quad a,b;$$
$$0,a;\quad 0,b;\quad 0,ab;\quad 0,ba;\quad b,a.$$

To these must be added the arrangement $(0,0)$ in which neither object is chosen. Thus the total is 11, as before.

Theorem VII. *The number of ways in which n different objects can be arranged in exactly r groups, not necessarily using all the objects, is the coefficient of x^{n-r} in the expansion of $n!\,e^x(1-x)^{-r}$.*

For we place one of the objects in each of the r groups, and we have then to distribute the $n-r$ remaining objects into r or fewer groups, as in the last theorem. Hence also, when the order of the groups among themselves is disregarded,

Theorem VIII. *The number of ways in which n different objects can be arranged in r indifferent groups, not necessarily using all the objects, is the coefficient of x^{n-r} in the expansion of $\frac{n!}{r!}e^x(1-x)^{-r}$.*

Suppose that we form n sets of letters from the set a_1, a_2, $a_3,\ldots,a_t$; suppose that the letter a_1 occurs in n_1 of the sets, that a_2 occurs in n_2 of the sets, while the number of sets containing a_1 and a_2 is n_{12}.

Then the number of sets containing a_1 only is n_1-n_{12}; the number containing a_2 only is n_2-n_{12}. Hence the number of sets containing either a_1 or a_2 only is $n_1+n_2-2n_{12}$, and the number containing at least one of a_1, a_2 is

$$n_1+n_2-2n_{12}+n_{12} = n_1+n_2-n_{12}.$$

It follows that the number of sets free from a_1 and a_2 is

$$n-(n_1+n_2)+n_{12}.$$

Let us consider now three letters a_1, a_2, a_3; suppose that n_3 of the sets contain a_3, that n_{23} contain a_2, a_3, that n_{31} contain a_3, a_1, while n_{123} contain a_1, a_2, a_3.

From the preceding result it follows that the number of sets containing at least one of a_2, a_3 is $n_2+n_3-n_{23}$; and the number containing at least one of a_1a_2, a_1a_3 is $n_{12}+n_{13}-n_{123}$. Hence the number of sets containing at least one of a_1, a_2, a_3 is

$$n_1+(n_2+n_3-n_{23})-(n_{12}+n_{13}-n_{123})$$
$$= n_1+n_2+n_3-(n_{12}+n_{23}+n_{31})+n_{123}.$$

Reasoning inductively in this manner we obtain the following general result:†

THEOREM IX. *If n sets of letters formed from $a_1, a_2, a_3, \ldots$ are such that the letter a_i occurs in n_i sets, the letters a_i, a_j occur in n_{ij} sets, the letters a_i, a_j, a_k occur in n_{ijk} sets, and so on, then the number of sets free from $a_1, a_2, \ldots, a_r$ is*

$$n-\sum_{i=1}^{r} n_i+\sum_{i,j=1}^{r} n_{ij}-\sum_{i,j,k=1}^{r} n_{ijk}+\ldots\pm n_{12\ldots r}.$$

COROLLARY 1. *If none of the sets is free from $a_1, a_2, \ldots, a_r$, then*

$$n-\sum n_i+\sum n_{ij}-\sum n_{ijk}+\ldots\pm n_{12\ldots r} = 0.$$

COROLLARY 2. By similar reasoning it may be shown that *the number of sets containing one only of the letters $a_1, a_2, a_3 \ldots$ is*

$$\sum n_i-2\sum n_{ij}+3\sum n_{ijk}-\ldots.$$

If $n_1 = n_2 = \ldots = N_1$ say, and $n_{12} = n_{23} = \ldots = N_2$, and so on, the number of sets free from the specified letters is

$$N-rN_1+\frac{r(r-1)}{2!}N_2-\frac{r(r-1)(r-2)}{3!}N_3+\ldots\pm N_r,$$

where r is the number of letters in question, and $N = n$.

EX. 1. If the n given sets are a_1, a_2, a_5, $a_1 a_2$, $a_2 a_3$, $a_1 a_4$, $a_4 a_5$, $a_1 a_2 a_3$, $a_1 a_2 a_5$, $a_3 a_4 a_5$, $a_1 a_3 a_4$, and the r specified letters are a_1, a_2, a_3, then $n = 11$, $r = 3$, $\sum n_i = 15$, $\sum n_{ij} = 7$, $n_{123} = 1$. Thus the number of sets free from a_1, a_2, a_3 is

$$11-15+7-1 = 2,$$

as is immediately verified.

EX. 2. At a school of 1,000 children, groups were examined for defective teeth, vision, and hearing, and the following results tabulated:

Numbers examined for:

Teeth .	180	Eyes and teeth .	90	Eyes, teeth, and hearing	40
Eyes .	700	Eyes and hearing .	170		
Hearing .	220	Teeth and hearing .	80		

The records of these cases were accidentally destroyed and it was not known how many of the children had actually been examined. What is the probability that a particular child was not examined?

By Theorem IX, the number of children not examined is

$$1{,}000-1{,}100+340-40 = 200.$$

Hence the required probability is $200/1{,}000 = 0{\cdot}2$.

† An equivalent theorem is given by Poincaré, *Calcul des Probabilités*; a particular form will be found in Whitworth, *Choice and Chance*, Chap. II.

Ex. 3. A certain factory produces and tests 7,000 motor-cars per year. The possible defects are catalogued as follows:

B = bodywork, C = chassis, E = engine, I = instruments.

Thus BCE denotes a case of compound defect in 'bodywork, chassis, and engine'. A year's record of defects is shown in the accompanying table:

$B = 120$	$BC = 50$	$BCE = 24$	$BCEI = 2$
$C = 150$	$BE = 40$	$BCI = 15$	
$E = 185$	$BI = 23$	$BEI = 5$	
$I = 200$	$CE = 55$	$CIE = 10$	
	$CI = 35$		
	$EI = 28$		

Find the percentage of cars which pass all four tests at the first trial.

Ex 4.† The number of ways in which a row of n objects can be deranged, so that no object remains in its proper place, is the greatest integer contained in $n!/e$.

For the total number of arrangements of the objects is $N = n!$. Of these, the number of arrangements in which at least one object is in its proper place is $N_1 = (n-1)!$; the number for which at least two objects are in their proper places is $N_2 = (n-2)!$, and so on.

Hence, by Theorem IX, the number of arrangements free from all these restrictions (i.e. for which all the objects are deranged) is

$$n!-\frac{n}{1!}(n-1)!+\frac{n(n-1)}{2!}(n-2)!-...$$

$$= n!\left(1-\frac{1}{1!}+\frac{1}{2!}-...\pm\frac{1}{n!}\right).$$

This number is certainly an integer; the last term is ± 1, so that if the series of terms in the brackets is replaced by e^{-1}, we merely add a fraction to the required number; whence the result.

Ex. 5. Two shuffled packs of 52 cards are dealt by two players, each dealing a card simultaneously. Show that the probability that all the 52 pairs of cards so dealt will be different is approximately $1/e$.

We may take one of the packs as specifying the order, which may be one of 52! arrangements. Then the number of ways in which the second pack may be arranged so that no card is in its proper place is $52!/e$, approximately. Hence the required probability is $1/e$, approximately.

The probability that identical cards will be dealt on at least one occasion is therefore $1-(1/e)$.

† This proposition is a variant of one due to Montmort (1708).

Ex. 6. A man writes a number (not less than nine) of letters and their corresponding envelopes. If the letters are inserted in the envelopes irrespective of the addresses, show that the probability that all the letters will go wrong is approximately $1/e$.

THEOREM X. *The number of ways in which n different objects can be arranged in r or fewer parcels is r^n.*

For each of the objects can be assigned to any one of the r parcels in r ways, and this gives r^n arrangements in all.

THEOREM XI. *The number of ways in which n different objects can be arranged in exactly r parcels is the coefficient of x^n in the expansion of $n!(e^x-1)^r$.*

For, by the last theorem, the number of arrangements in which blanks are admissible is r^n. The number of arrangements in which one assigned blank is admissible is $(r-1)^n$, and so on. Hence, by Theorem IX, the number of arrangements in which no blanks are admissible is

$$r^n-\frac{r}{1!}(r-1)^n+\frac{r(r-1)}{2!}(r-2)^n-\ldots\pm\frac{r(r-1)}{2!}2^n\mp r.$$

Now $$(e^x-1)^r = e^{rx}-\frac{re^{(r-1)x}}{1!}+\frac{r(r-1)}{2!}e^{(r-2)x}-\ldots.$$

Hence, by the exponential theorem, the coefficient of x^n in this expansion is

$$\frac{r^n}{n!}-\frac{r}{1!}\,\frac{(r-1)^n}{n!}+\frac{r(r-1)}{2!}\,\frac{(r-2)^n}{n!}-\ldots,$$

whence the above result.

THEOREM XII. *The number of ways in which n identical objects can be distributed into r parcels such that no parcel contains less than q objects or more than $q+t-1$, is the coefficient of x^{n-qr} in the expansion of $(1-x^t)^r(1-x)^{-r}$.*

It is clear that the required number is the coefficient of x^n in the product of the r factors

$$(x^q+x^{q+1}+\ldots+x^{q+t-1})(x^q+x^{q+1}+\ldots+x^{q+t-1})\ldots,$$

that is, in $$x^{qr}(1+x+x^2+\ldots+x^{t-1})^r,$$

or in $$x^{qr}(1-x^t)^r/(1-x)^r.$$

Hence the number sought is the coefficient of x^{n-qr} in $(1-x^t)^r(1-x)^{-r}$.

Ex. 1. A die whose faces are numbered from 1 to 6 is thrown four times; in how many ways can the number 8 be obtained in the four throws?

In this case we require the coefficient of x^8 in the product

$$(x^1+x^2+\dots+x^6)^4,$$

i.e. the coefficient of x^4 in the product $(1+x+x^2+\dots+x^5)^4$.

To find this coefficient we write the latter expression as $(1-x^6)^4/(1-x)^4$ and, supposing that $x < 1$, we expand $(1-x)^{-4}$ as a binomial series. Thus we require the coefficient of x^4 in

$$(1-4x^6+\dots)(1+4x+10x^2+20x^3+35x^4+\dots),$$

i.e. 35.

Note that the total number of possible combinations of the numbers 1 to 6, in four throws, is the sum of all the coefficients in $(x+x^2+\dots+x^6)^4$, and this is obtained by putting $x = 1$; the number is therefore 6^4.

Ex. 2. The probability that a die which is thrown four times gives a total of 8 is $\frac{35}{6^4} = \frac{35}{36^2} = \frac{1}{36}$, approximately.

Ex. 3. Show that the probability that the number m will be obtained by throwing a die r times is the coefficient of x^m in the expansion of

$$x^r(1-x^6)^r(1-x)^{-r}/6^r.$$

Ex. 4. Given the two sets of numbers 1, 2, 3, 4, 5; 1, 3, 5, 7, 9, find the probability that the sum of two numbers selected, one from each group, is 8.

The number of possible pairs of numbers is $5^2 = 25$; of these the number of pairs whose sum is 8 is evidently 3; thus the probability is 3/25.

Ex. 5. Given the three sets of numbers 1, 2, 3, 4, 5; 1, 3, 5, 7, 9; 2, 4, 6, 8, 10, find the probability that the sum of three numbers selected, one from each set, should be 16.

The number of sets whose sum is 16 is the coefficient of x^{16} in the product $(x+x^2+\dots+x^5)(x+x^3+\dots+x^9)(x^2+x^4+\dots+x^{10})$, i.e. the coefficient of x^{12} in $(1+x+\dots+x^4)(1+x^2+\dots+x^8)^2$, which is 12.

Hence the probability is 12/125.

Ex. 6. A set of 10 cards is marked with the numbers 2, 4,..., 20. In how many ways can a total of 36 be found in a hand of 4 cards?

EXAMPLES ON CHAPTER VII

Ex. 1. Four men arrange to meet at the 'White Hart' tavern in a certain town. It happens that there are four taverns with that name; show that the probability that all the men choose different taverns is $\frac{3}{32}$.

Ex. 2. If n people seat themselves at a round table, show that the probability that two individuals are neighbours is $2/(n-1)$.

Ex. 3. A pack of 52 cards is dealt out to four players; show, by Stirling's Theorem, that the probability that the whole of one particular suit is dealt to one particular player is approximately $156/10^{14}$.

Ex. 4. Show that the probability of obtaining 14 is the same with 3 dice as with 5.

Ex. 5. A die is thrown 10 times; prove that the probability that every face appears at least once is 38,045/139,968.

Ex. 6. A set of r consecutive numbers is selected from the numbers $1, 2, \ldots, n$; if a second set of s consecutive numbers is selected, what is the probability that it has no number in common with the first?

Ex. 7. Find the probability of throwing not more than 8 with 3 dice.

Ex. 8. Show that there is a greater probability of obtaining 9 in a single throw with 3 dice than with 2.

Ex. 9. There are n houses in each of which the population may vary from 1 to n. What is the probability that the average population per house is 4?

Ex. 10. Show that the most probable sum to be obtained by throwing $2n$ dice is $7n$, and that with $2n+1$ dice both $7n+3$ and $7n+4$ are equally likely.

Ex. 11. Find the number of positive integral solutions of the equation

$$x+y+z+u = 12,$$

if the unknowns are to lie between 1 and 6.

Ex. 12. Given m kinds of objects and n of each kind, show that the probability that $m-r$ selected objects will be all different is

$$^{mn+r}C_r\, n^{m-r}/^{mn+r}C_m.$$

Ex. 13. If a coin is tossed $2n$ times, prove that

(i) the probability that the numbers of heads and tails obtained are equal for the first time at the $2n$th throw is $^{2n}C_n/4^n(2n-1)$;

(ii) the probability that in $2n$ throws the numbers of heads and tails are never equal is $^{2n}C_n 4^n$.

(iii) the probability that the numbers of heads and tails have been equal once and only once is $^{2n}C_n/4^n$.

Ex. 14. Prove that the number of ways of obtaining the sum r with n dice is

$$^{r-1}C_{n-1} - {^nC_1}\, {^{r-7}C_{n-1}} + {^nC_2}\, {^{r-13}C_{n-1}} \cdots.$$

Ex. 15. If a coin is tossed n times, show that the probability that there will not be a consecutive heads is the coefficient of x^n in the expansion of $\dfrac{1}{2^n}\,\dfrac{1+x+x^2+\ldots+x^{a-1}}{1-x-x^2-\ldots-x^a}$.

Ex. 16. If m objects be distributed among a men and b women, show that the probability that the number received by the men is odd, is

$$\{\tfrac{1}{2}(b+a)^m - \tfrac{1}{2}(b-a)^m\}/(b+a)^m \qquad (b > a).$$

Ex. 17. Among a batch of 240 eggs, 12 are bad. The eggs are sent in cartons of a dozen to 20 different customers. Find the probability that

(i) a particular customer will receive two or more bad eggs,

(ii) two particular customers will receive two or more bad eggs,

(iii) all the bad eggs are delivered to three customers.

Ex. 18. In a street of 100 houses 25 are known to have defective drains, 75 have broken windows, and 15 have both defective drains and broken windows. Show that the probability that a given house is sound in windows and drains is 3/20.

Ex. 19. D_1 and D_2 are two diseases such that the probability of any one infected with D_1 acquiring D_2 from an infected individual is p_1, and the probability of any one infected with D_2 acquiring D_1 from an infected individual is p_2. Suppose that the diseases cannot be acquired save by mutual contagion, and that n_1 and n_2 people infected with D_1 and D_2 respectively come to live in a town of n inhabitants, mixing freely with them. What is the probability that an inhabitant will be free from both or either of the diseases?

Ex. 20. A billposter has 100 placards to post in sets of 3 or 4. If the placards contain 10 different types of 10 each, find the probability that a given set of 3 will have 2 alike.

CHAPTER VIII

THE EMPIRICAL THEORY OF DISTRIBUTIONS

1. Hypothetical populations and typical constants

So far we have been concerned with probability as a mathematical subject of study, the category (1) of Chapter II. In this section we turn to the consideration of category (2), which concerns itself in the first place with enumerating the frequency of occurrence of actual events in a physical problem. Once again let us emphasize the difference between (1) and (2): in the analysis so far developed (1) has dealt with the enumeration of all possible arrangements that can be conceived to occur in any given situation; on the other hand, (2) is concerned with the actual events as they have occurred in circumstances akin to those in which the results are to be applied. The crucial question which has to be faced, in the use of mathematical probability in the theory of statistics, is how the mathematical theorems of (1) can legitimately be combined with the empirical data of (2) to enable predictions to be made about forthcoming events of the type (2).

We begin with a discussion of *Histograms*, a pictorial arrangement of physical data in a form suitable for mathematical analysis.

Let us suppose that 100 leaves are stripped from a tree and their mean widths measured; it is then found that these lie between 1 and 2 inches in the proportions shown in the table. To represent our data graphically we set off unit length on the

Width in inches	*No. of leaves*
1·0 to 1·1	8
1·1 to 1·2	10
1·2 to 1·3	15
1·3 to 1·4	20
1·4 to 1·5	18
1·5 to 1·6	11
1·6 to 1·7	7
1·7 to 1·8	6
1·8 to 1·9	3
1·9 to 2·0	2

x-axis, divided into tenths, and at the mid-point of each interval we erect an ordinate proportional to the number of leaves to be found in that interval. By drawing a system of horizontal and vertical lines as shown, we obtain a step-curve, called a 'histogram'.

It is clear that, by reducing the ordinates in a certain ratio, the histogram can immediately be converted into a mathematical probability diagram; since there are 100 members of the population considered, the proportions belonging to the subclasses (1, 1·1), (1·1, 1·2),... are respectively 8/100, 10/100, etc. These proportions represent, in the mathematical sense, the probability of occurrence of the subclasses among the population of 100 leaves.

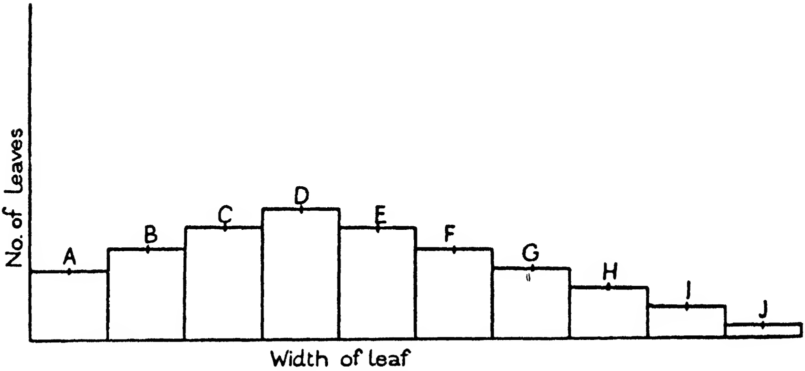


Fig. 19

Once more we stress the distinction between mathematical and empirical probability by asking two questions:

(i) What is the mathematical probability that a leaf known to be a member of this population of 100 leaves has a width lying between 1·2 and 1·3 inches? The answer is 15/100.

(ii) What is the 'probability' that yet another leaf known to have been stripped from the tree containing the original 100 has a width lying between 1·2 and 1·3 inches? So far we have attached no significance whatever to this interpretation of probability. Before any step can be taken enabling us to give a sensible answer to this question, we require some information concerning the nature of the larger population from which the population of 100 leaves has been drawn, or—as is sometimes stated—we require to know whether the latter is a 'fair sample' of the original population. The answer to the question, therefore, cannot be divorced from the assumed criterion of the 'fairness' of the sample.

Probability Curves

The simple laws of mathematical probability given in Chapter IV can be illustrated from the above diagram. For example, the probability that a leaf defined as a member of the population of 100 has a width lying between 1·1 and 1·4 inches is $\frac{10+15+20}{100}$ = sum of the probabilities that its width lies in the ranges (1·1, 1·2), (1·2, 1·3), (1·3, 1·4). The probability that the width lies somewhere in the range (1, 2) is obviously unity. The probability that the leaf has a width lying in the range (1·1, 1·4) is the area between the probability diagram, the x-axis, and the ordinates at 1·1, 1·4.

Frequency and Probability Curves

If through $ABC...J$ we draw a continuous curve such that the area under each element of curve is equal to the area of the corresponding rectangle in the histogram, the curve so obtained is called the 'frequency curve'; if the ordinates of this curve be reduced in the ratio 1 : 100, as in the formation of the probability diagram, we derive a *probability curve.* For this curve also we can state that the probability of a leaf having a width lying in the range (1·2, 1·6), say, is measured by the area under the curve; that is, if $y = p(x)$ is the equation of the curve, the required probability is

$$P = \int_{1\cdot 2}^{1\cdot 6} p(x)\,dx.$$

It should be noted that we are not justified in stating that the probability that a leaf has a width lying in the range (1·23, 1·63) is $\int_{1\cdot 23}^{1\cdot 63} p(x)\,dx$. It may be convenient (as we shall find) for the purpose of mathematical treatment to assume that the probability of an individual specimen having a width lying in the range (a, b) is $P = \int_a^b p(x)\,dx$; but we should have to justify such an assumption or, alternatively, to find some measure for the extent of the error involved in making it.

Ex. 1. In the examination of 148 pods of large yellow broom, the frequency of seeds in a pod was found to be as follows:

No. of seeds	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
No. of pods	0	0	1	2	6	12	12	7	7	14	16	16	14	11	9	8	6	4	2	1

Construct the histogram and the frequency curve for this population.

Ex. 2. A second batch of such pods was measured and the frequency of their lengths obtained, as follows:

Length	*Frequency*
2·2–2·8	0
2·8–3·4	1
3·4–4·0	3
4·0–4·6	20·5
4·6–5·2	11
5·2–5·8	23·5
5·8–6·4	10·5
6·4–7·0	3·5

Construct the frequency curve.

Probability as a Continuous Function

If we are to justify the above-mentioned assumption that a probability may be regarded as a continuous function of a variable in experimental practice, we are faced with what at first appears to be a difficult problem concerning the continuity of natural phenomena. We have remarked that all observations are obtained, at some stage or other, by the use of a measuring scale; and if the process of measurement is examined, it is found to consist in an attempt to make two marks on the scale coincide with two marks on the object measured. But whereas it is possible to make one mark on the scale coincide, to our satisfaction, with a mark on the object, the other mark in general falls somewhere between two adjacent marks on the scale. Even when the accuracy of the measurement is increased by the use of a vernier, say, invariably the reading of the scale division involves an estimate which is equivalent to stating that the mark does not fall between two scale divisions, but on one or other of them. There always exists a finite 'jump' corresponding to the least interval which can be measured by means of the scale.

The same kind of restriction is implicit in any tabulated set of numbers, such as a table of logarithms or trigonometric functions; in fact, by no set of numbers or measurements can we represent fully a continuous function. Two leaves out of a batch of 10,000 will be classed as of equal width if with our measuring rod we cannot detect any difference in their widths; nevertheless the difference, if any, between two widths, that might be detected by a more accurate process, may correspond to a finite jump which we ignore in the measurement. While, therefore, it is clear that all measurements obtained from Nature must show discontinuity and all frequency curves constructed from them ought strictly to be histograms, it would be unreasonable to assert that for our purpose we must regard the growth of leaves, say, necessarily as a discontinuous process. An apparent discontinuity arises from limitations in our method of measurement, but it is unnecessary to import these into our analysis. From our standpoint the distinction between continuity and discontinuity in these cases amounts to little more than stating that we take the area between the histogram and the x-axis to be equivalent to the area under a continuous curve passing through the vertices of the histogram, it being supposed that the error so committed is small. If it is a great convenience for us to deal with a continuous curve rather than with a histogram, the loss in accuracy, even if it were perceptible, would be more than compensated for by the gain in power.

The Meaning of 'Population'

Here the empirical data have been used for constructing a histogram which in its continuous form represents the mathematical probability curve. In passing from the former to the latter we are in effect constructing a hypothetical population on the basis of the experimental sample. It is customary to represent such a continuous curve in mathematical form and then to assume, either explicitly or implicitly, that the form so obtained has a validity for a range of the variable much beyond that found in the given sample. This process is tantamount to extrapolating the population by means of a mathematical expression.

In discussing the validity of an application of mathematical probability or statistical theory to scientific experiment, there are several questions that merit examination. Let us contrast, in the first instance, the conduct of a physical experiment with the collection of botanical data, e.g. for determining the size of leaves on a particular type of tree. In his experiment the physicist is able to exercise a considerable degree of control over the situation; he can plan and lay out the environment; he can, in general, eliminate what are called 'systematic errors' or even periodic fluctuations. The consequences are twofold. In the first place he can state from the beginning that the quantity he is measuring will lie within a prescribed and comparatively narrow range; he will know, for example, that the expansion of a metal rod in certain circumstances cannot be more than 0·5 cm. or less than 0·2 cm. This he knows on the basis of his past experience of scientific inquiry, and it would be extraordinarily rare for an experiment to be conducted without some such preliminary knowledge.† In the second place, the actual experiment which he performs narrows this range still further; the observations obtained show that the 'true readings' are grouped within a much smaller band of values. Moreover, because of the fact that the experiment has been carefully performed and the measurements made *after* a series of delicate adjustments, the scientist is perfectly well aware that to multiply the number of readings merely to satisfy the demands of the statistician cannot possibly increase his accuracy—they may succeed only in encouraging him to incorporate a number of less accurate observations in his results.

When we consider the collection of botanical data the conditions are seen to be very different. The botanist has to take the material with which Nature provides him, largely in circumstances over which he has no control. His data may therefore range over wide regions; he can, like the physicist, state in advance upper and lower limits within which his measurements will lie, but the narrow band will be much less accurately defined: the more observations he can collect, the greater will be his knowledge of the features he is studying. Whereas the physicist can proceed on the experimental assumption that there is a definite expansion of the rod to which his measurements are approximations, the botanist cannot assert that there

† The far-reaching effects of any exception to this rule can be seen from the consequences of the Michelson-Morley experiment.

is a definite size of leaf, the 'true' size, to which his collection approximates. One of the purposes of his experiment is in fact to discover whether he can usefully apply such a fiction to his subject-matter.

In the light of the above facts concerning experimental practice in physics, it must be admitted that in many cases there is no justification for the assertion that the limited set of data obtained by an experimenter are a sample of a hypothetical population or a much wider collection.[†] The position is different when we are dealing with biological phenomena of the type mentioned, for here the actual collection of data has to be seen as a step towards the building up of the hypothetical population, with its special conception of a 'true' value. This makes the application of statistical theory to physical experiment a much more delicate and uncertain procedure than to biological, meteorological, or economic phenomena.

The type of collection or hypothetical population which we have had in mind is a static unchanging one. But such is by no means the only possible type. In the paper referred to above, Campbell illustrates the difficulty of assigning two different samples to different collections by considering the rainfall records of 1901–20. 'Was the climate between 1901 and 1910', he asks, 'different from that between 1911 and 1920? If this problem is statistical, the records for 1901–10 and for 1911–20 must be samples of two possibly different collections. But what are the remainders of these collections? Not the records for other years; for, if the climate may be changing, other years are not comparable. But meteorological records must be records for some defined period. If the records for 1901–10 are a mere sample of the records for some longer period, and not the whole collection relevant to the problem, what is this longer period?'

The answer to these conundrums surely lies in the fact that the climate of a country is itself a varying phenomenon and therefore the two records for 1901–10 and 1911–20 must be regarded as successive samples of a *varying* hypothetical population. Whether these samples provide data adequate for the drawing of valid conclusions about climatic changes as a whole is another matter. All we wish to point out is, that unless the

[†] Cf. N. Campbell, *Proc. Phys. Soc.* **47** (1935), 800.

records in question be regarded as successive samples of a *varying* population, inconsistences of the type indicated by Campbell are bound to arise.

But we must not over-estimate the importance of such matters in experimental practice; we shall certainly do so if we imagine that all experiment is necessarily individual. When scientific method demands that a particular conclusion shall be accepted only if it is accorded general assent, this should mean not only that the experiment which led to it is 'accepted' as from one research worker and that it can be imagined repeated if necessary, but that it is in fact repeated by a number of other workers. Thus, many measurements have been made of the velocity of light, by different observers working under diverse conditions or by the same observer using a variety of methods. For the final conclusion to be acceptable, the collection of data has to be regarded as a 'fair sample' of what scientists who perform the experiment are likely to find. On the other hand, the search for a true scientific entity would be fruitless unless all the numbers obtained could be regarded as clustering about some so-called 'true value'. The set of observations so found therefore embody a series of diverse conditions of experiment which are necessarily unspecifiable in detail; and in essential contrast with the case of the individual experimenter, the larger the amount of such observations, the greater the precision with which the true value can be stated. For this reason it is of vital importance that the mass of data found by different observers should form a coherent collection; they have to be unified, and the unifying process which attempts to cancel out the numerous irrelevant circumstances is essentially a statistical one. As we have remarked, each experimenter will be able to state at the beginning that the quantity he proposes to measure will lie within a prescribed, comparatively narrow range; the fact that this range is practically identical for all the observers is merely evidence that they all begin with the same basic knowledge of the problem. The narrower range which emerges in each experiment will reflect among other things the diverse conditions of the individual experiment, and it is these ranges that have to be dealt with in a statistical manner. In disagreement, therefore, with the point of view put by Campbell,† we hold that a statistical approach to observational data derived from different observers (or from the same observer working under different conditions) is inescapable and is in fact fundamental in the development of science itself.

Following up this idea, we shall seek to discover what are the most suitable probability functions which can be utilized in practical cases, as they occur. We are then justified in assuming that the probability of occurrence of a variable in the range (α, β) is not only to be obtained by computing the area between the histogram and the x-axis, for that range, but by evaluating $\int_{\alpha}^{\beta} p(x)\,dx$, where $y = p(x)$ is now a continuous curve passing through or near the vertices of the histogram.

Typical Constants

For experimental purposes, and particularly for the construction of hypothetical populations, it is inconvenient to handle a mass of detailed data. It is therefore necessary to examine whether certain characteristics of the data may suffice for the purpose in view. We pose the general problem as follows: Given a set of numbers $a_1, a_2,..., a_n$, can we find a single number which can be regarded as a measure typical of the set? Thus, $a_1, a_2,...$ may be the numbers obtained in measuring a desk (as in Chapter II), and we may inquire, can we find a single number which can be regarded as typical and which can be referred to, for our purposes, as *the* length?

If we desire to specify the set $a_1, a_2,...$ even more precisely than is possible by using a single number, a second problem arises, namely, how closely are the members of the set packed or distributed about the 'typical' member? We shall, of course, have to make precise the meaning of the word 'typical' in the given context. That this second problem is closely connected with the concept of frequency is seen if we state it in this way: How frequently do the measured members of the set fall into the successive ranges of, say, 0·1, measured from a 'typical' member? These two questions require to be answered very precisely before further steps can be taken to handle a set of

† Loc. cit. 808.

data adequately in terms of what may be called its typical constants.

What characteristic shall we expect our first typical constant to possess? If it were a large positive number, the differences between it and the actual readings would be large also; similarly if it were large and negative. There should be a typical constant lying somewhere between these two extremes, such that the sum of the differences, taken positively, has a smallest value; it would be a number about which the set as a whole is most closely packed, in accordance with the requirements we have already indicated. This suggests either that the sum of the absolute values of the differences between it and the actual readings should be a minimum, or that the sum of the even powers of these differences should be a minimum. Each of these suggestions would give us a typical constant upon which to base our discussion.

Let us illustrate by a problem. Consider the set of numbers 2, 7, 5, 15, 10, 4; take any number x and write down the differences $x-2$, $x-7$, etc., some of which may be positive and some negative. The sum of the squares of these differences is

$$(x-2)^2+(x-7)^2+...+(x-4)^2 = y, \text{ say.}$$

If we plot the values of y against x we obtain a parabola whose minimum ordinate occurs at $x = \frac{2+7+...+4}{6}$ = the average of the given numbers.

This minimum ordinate thus represents the least value of the sum of the squares of the deviations of x from the given numbers; it is attained when x has the 'average value'. If we define the typical constant in this case as that value of x which makes the sum of the squares a minimum, then we find it by taking the average of the given numbers.

The proposition is true in general. Thus, let a_1, a_2,..., a_n be a set of numbers, of which x is the typical value. The sum of the squares of the deviations is

$$y = (x-a_1)^2+(x-a_2)^2+...+(x-a_n)^2.$$

This attains its minimum value when $dy/dx = 0$, i.e. when

$$(x-a_1)+(x-a_2)+...+(x-a_n) = 0,$$

so that the required value of x is $(a_1+a_2+...+a_n)/n$, the average value.

The minimum value divided by n is called the square of the *standard deviation* σ: or if a is the average value of a_1, a_2,..., a_n, we have

$$\sigma = \surd\{(a-a_1)^2+(a-a_2)^2+...+(a-a_n)^2\}/\surd n.$$

[*Note upon 'average' and 'mean'*

If a train travelling between two stations changes its speed steadily from 40 to 50 miles per hour, its average speed is 45 miles per hour. If the passengers in the train have heights varying from 5 ft. 5 in. to 6 ft. 1 in., they may have an average height of, say, 5 ft. 8 in. In the first case it is legitimate to assume that at some point in the journey the train has actually been travelling at 45 miles per hour; in the second it does not follow that any one of the passengers has a height of 5 ft. 8 in. —if we refer to it as a height, it is a fictitious one.

It is usual to apply the terms 'average' and (arithmetic) 'mean' indiscriminately to these two cases; but since a real distinction exists between them it would perhaps be worth while, for the sake of clarity, to say that the mean speed of the train is 45 miles per hour, while the average height of the passengers is 5 ft. 8 in. A member of the class would then occupy the position of the mean, but there need be no member of the class which possesses the 'average' characteristic.]

We remark that σ and a are both 'typical' constants, although the former has been found in the attempt to discover the latter. Since σ^2 is the mean value of the squares of the deviations of each member of a_1, a_2,..., a_n from its average, σ (the 'root mean square') gives us an overall measure of the deviation of the set from the average a, without reference to sign.

There are two other features of the set which are sometimes found useful. Suppose that a frequency diagram has been constructed in which the ordinates represent the number of readings lying in successive intervals. The interval in which the ordinate attains its maximum clearly corresponds to the most frequent or 'most fashionable' value of x among the set. This value is called the *mode*; in general it is not identical with the average or mean, but it will be if the frequency curve is symmetrical about the mean value. A frequency curve may have more than one mode; but we are here concerned only with cases in which a single mode exists.

Again, we may arrange our data in ascending order of magnitude and divide them into two sections half-way, so that as many measurements lie above this division as below it. This position is called the *median* and is such that the probability of any member of the set lying below (or above) it is $\frac{1}{2}$.

Measure of the significance of σ

The magnitude of σ alone may not, of course, provide us with all the information we may desire, even when it is associated with the average value a. As a next step we may inquire how many of our readings lie within the range $(-\sigma, \sigma)$ about a, and how many outside; or, as we may ask, what is the probability that a member of the set deviates from a by more than σ?

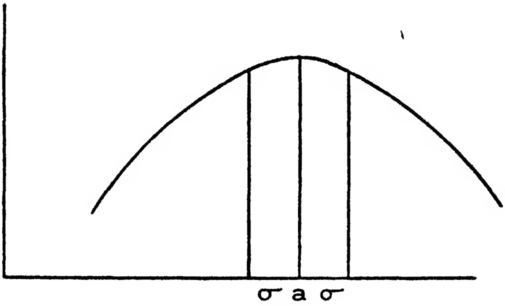


Fig. 20

The answer to this question may be found at once from a knowledge of the average, the standard deviation, and a histogram or a graph of the frequency curve. In the frequency diagram we erect ordinates at the points $x = a$, $x = a \pm \sigma$; the number of observations which fall within the range indicated, divided by the total number, is a measure of the probability of the subclass whose deviations from the average are less in absolute value than the standard deviation. In accordance, therefore, with our definition, this determines the probability that any individual observation, as a member of the hypothetical population specified by the continuous curve, has a deviation less than σ; if this probability is 'high', the set is 'closely packed' about the average. We note that 'high' is here a matter of judgement.

If $p(x)$ is the probability function in the given case, then the required probability is evidently $\int_{a-\sigma}^{a+\sigma} p(x)\,dx$. It is usual to take the origin of coordinates at $x = a$, since many frequency curves are symmetrical about the ordinate erected there. If $P(x)$ is the transformed probability function, the probability is now

$$\int_{-\sigma}^{\sigma} P(x)\,dx.$$

An alternative constant associated with the distribution is suggested by the question: for what deviation from the average is it equally probable that an observation will fall within, as without, the range? Analytically, we inquire for what deviation λ is

$$\int_{-\lambda}^{\lambda} P(x)\,dx = \tfrac{1}{2}.$$

In any given case, the value of λ can be determined by actual enumeration or, if the hypothetical frequency curve has been constructed, by any method for evaluating areas. The value of λ so defined is called the 'probable error' (a misnomer if by that term we are led to conceive of it as the most probable error). If a deviation from the average is indeed to be regarded as an 'error', as though the average were the 'truth',† then every error has its appropriate probability. In the case where the deviation is $\pm\sigma$ we take the probability to measure the extent to which the observations are packed about the average; in the case where the error is λ, the probability is $\frac{1}{2}$.

Thus we have been led to a succession of typical constants in the attempt to specify a distribution. These are

(1) the average a;

(2) the standard deviation σ about the average;

(3) the probability p that an observation has a deviation from the average of less than the standard deviation;

(4) the probable error.

Which of the constants will suffice in any given case depends on their magnitude, our judgement of what their magnitude

† See note on 'average' and 'mean', p. 112.

implies, and the purpose for which the data are to be used. If the standard deviation is small, then the average itself may suffice; if the probability p is great (i.e. in the neighbourhood of 1), then a and σ may suffice. If not, the 'probable error' gives us some further indication of the extent to which the distribution curve is dispersed about the average. We shall analyse these circumstances in greater detail when we come to study particular forms of probability curves.

Definition of Weights

If $x_1, x_2,..., x_n$ are a set of observations such that x_1 occurs p_1 times, x_2 occurs p_2 times, ..., and x_n, p_n times, then the total number of observations present is

$$p_1+p_2+...+p_n = \sum p.$$

The sum of the observations is

$$p_1x_1+p_2x_2+...+p_nx_n = \sum px.$$

Thus the average $a = \sum px/\sum p$.

The numbers $p_1, p_2,..., p_n$ are called the *weights* of the observations $x_1, x_2,..., x_n$. It is clear from the formula that all the weights may be multiplied by the same arbitrary constant without affecting the value of the average. If at points whose abscissae are $x = x_1, x_2,..., x_n$, ordinates of lengths $p_1, p_2,..., p_n$ are erected, the diagram so obtained is a histogram, as we have already seen.

Typical constants for a continuous distribution

All the constants so far defined are relevant to actual experimental data. We have seen that when we proceed to replace the histogram by a continuous probability curve we are in effect postulating a hypothetical population. We proceed now, therefore, to the derivation of analogous constants for the latter. Let the equation of the hypothetical probability curve be $y = p(x)$; we now seek a typical constant a by forming the sum of the squares of the deviations of x from this constant. Thus, since a deviation $x-a$ in the interval dx occurs $p(x)$ times, the sum of the squares of the deviations is represented by $I \equiv \int\limits_{\alpha}^{\beta} (x-a)^2p(x)\,dx$, where α and β specify the range of the probability curve. We wish to find the value of a, if any, for which this integral is a minimum. We have

$$I \equiv \int\limits_{\alpha}^{\beta} (x-a)^2 p(x)\,dx$$

$$= \int\limits_{\alpha}^{\beta} x^2 p(x)\,dx - 2a\int\limits_{\alpha}^{\beta} xp(x)\,dx + a^2\int\limits_{\alpha}^{\beta} p(x)\,dx.$$

Now I will be a maximum or minimum with respect to a if $dI/da = 0$, that is, if

$$-2\int\limits_{\alpha}^{\beta} xp(x)\,dx + 2a\int\limits_{\alpha}^{\beta} p(x)\,dx = 0.$$

Thus $$a = \int\limits_{\alpha}^{\beta} xp(x)\,dx \bigg/ \int\limits_{\alpha}^{\beta} p(x)\,dx,$$

giving a value for a which obviously corresponds to the average of a set of observations when the number of such observations is finite.

That the value of a so found makes I a minimum follows from the fact that $\dfrac{d^2I}{da^2} = 2\int\limits_{\alpha}^{\beta} p(x)\,dx$, which is necessarily positive since $p(x)$ is everywhere positive.

We have thus obtained an extended concept of an average; by analogy, the standard deviation σ for the hypothetical population is defined by the relation

$$\sigma^2 = \int\limits_{\alpha}^{\beta} (x-a)^2 p(x)\,dx \bigg/ \int\limits_{\alpha}^{\beta} p(x)\,dx$$

$$= \left\{\int\limits_{\alpha}^{\beta} x^2 p(x)\,dx - 2a\int\limits_{\alpha}^{\beta} xp(x)\,dx + a^2\int\limits_{\alpha}^{\beta} p(x)\,dx\right\} \bigg/ \int\limits_{\alpha}^{\beta} p(x)\,dx$$

$$= \int\limits_{\alpha}^{\beta} x^2 p(x)\,dx \bigg/ \int\limits_{\alpha}^{\beta} p(x)\,dx - \left(\int\limits_{\alpha}^{\beta} xp(x)\,dx \bigg/ \int\limits_{\alpha}^{\beta} p(x)\,dx\right)^2,$$

in virtue of the expression found for a.

By calculating σ in any given case we can once again estimate the probability that any number in the range (α, β) will differ from the average by less than σ.

Ex. Suppose that a hypothetical population ranges in magnitude from 0 to 2 with a frequency which between 0 and 1 is given by $p(x) = x$, and between 1 and 2 by $p(x) = 2-x$.

Then $$\int_0^2 xp(x)\,dx = \int_0^1 x^2\,dx + \int_1^2 x(2-x)\,dx = 1.$$

Also $$\int_0^2 p(x)\,dx = 1.$$

Hence the average $a = 1$.

The standard deviation σ is given by

$$\sigma^2 = \int_0^2 x^2p(x)\,dx \Big/ \int_0^2 p(x)\,dx - a^2 = \tfrac{1}{6}.$$

Hence $$\sigma = \frac{1}{\surd 6}.$$

The probability that a member of the set between 0 and 2 will differ from the average by an amount $1/\surd 6$ is

$$\int_{a-\sigma}^{a+\sigma} p(x)\,dx = \int_{1-\frac{1}{\surd 6}}^{1} x\,dx + \int_{1}^{1+\frac{1}{\surd 6}} (2-x)\,dx = \frac{2\surd 6-1}{6}.$$

Tchebycheff's Theorem

Let $x_1, x_2,..., x_n$ be a set of n numbers; their mean $\bar{x}$ and standard deviation σ are then given by

$$n\bar{x} = \sum_1^n x_n, \tag{1}$$

$$n\sigma^2 = \sum_1^n (x_n-\bar{x})^2. \tag{2}$$

If λ is any positive proper fraction it follows that not more than $\lambda^2 n$ of the x's can deviate by more than σ/λ from $\bar{x}$. For suppose that $\lambda^2 n$ of them deviated to this extent at least from $\bar{x}$; then the sum of the squares of their deviations would exceed

$$\lambda^2 n\left(\frac{\sigma}{\lambda}\right)^2 = n\sigma^2,$$

which is a contradiction of (2).

It follows that whatever the nature of the distribution, the proportion of x's deviating from the mean by more than

2σ is less than $\frac{1}{4}$,

3σ ,, ,, $\frac{1}{9}$,

4σ ,, ,, $\frac{1}{16}$.

These figures provide an upper limit to the probability that a member of a set deviates from the mean by more than a given multiple of the standard deviation. For any given distribution this probability is, of course, easily calculated.

2. The Gaussian Law

In discussing the specification of typical constants we proceeded from the assumption, unwarranted except for its general plausibility, that one of these constants is such that the sum of the squares of the deviations of observations from it should be a minimum. We propose in this section to carry the problem of typical constants a stage further, by an elementary study of a number of hypothetical populations.

Let $y = \phi(x)$ be the equation to a probability curve giving the probability of an observation of measure x; we shall suppose that all the measurements which might be made in the given case, by a particular process, conform to this law of probability. Let $x_1, x_2, \ldots, x_n$ be a set of n of these measurements. Suppose that we move the origin of coordinates to a point on the x-axis, distant a from the present origin, where a is to be specified; the equation of the probability curve is now $y = \phi(\xi)$, where $\xi = x-a$, and the deviations of the given observations from a are

$$\xi_1 = x_1-a,\ \xi_2 = x_2-a,\ \ldots,\ \xi_n = x_n-a.$$

The probabilities that deviations $\xi_1, \xi_2, \ldots, \xi_n$ will separately occur are $\phi(\xi_1), \phi(\xi_2), \ldots, \phi(\xi_n)$. Hence the compound probability that out of all the possible deviations that might occur when n observations are made, precisely this combination arises, is the product

$$P = \phi(\xi_1)\phi(\xi_2)\ldots\phi(\xi_n).$$

We shall define the typical constant a to be such as to make the probability of precisely $\phi(\xi_1)\phi(\xi_2)\ldots\phi(\xi_n)$ occurring, greater than that for any other value of the constant.† That P attains its greatest value for some value of a does not necessarily mean that it attains a mathematical maximum, if we restrict a to lie in the range of the observations $x_1, x_2, \ldots, x_n$. We may suppose that a and $\xi_1, \xi_2, \ldots, \xi_n$ vary continuously over this range, but even then P is not necessarily a continuous function of a.

† This principle, in extended form, is applied in Chapter IX for the determination of hypothetical populations in general.

(We have already seen that a probability curve, for a set of given observations, is not necessarily continuous.) Accordingly we make the following additional assumptions:

(1) P, regarded as a function of $\xi_1, \xi_2, \ldots, \xi_n$ and a, is continuous and differentiable over the whole range.

(2) There is a single greatest value of P in the range which is also a maximum of the function.

(3) The value of a which makes P a maximum is the average value already determined.

We shall have to examine whether these assumptions can be fulfilled; certainly they impose restrictions on the nature of the probability function which will be reflected in the form eventually found for it. Whether they are such as to make the results inapplicable in practice is at the moment an open question.

Differentiating the function P logarithmically, we obtain the first condition for a maximum,

$$\frac{\phi'(\xi_1)}{\phi(\xi_1)}\frac{d\xi_1}{da}+\ldots+\frac{\phi'(\xi_n)}{\phi(\xi_n)}\frac{d\xi_n}{da}=0. \qquad (1)$$

But since $\xi_1 = x_1-a$, $\xi_2 = x_2-a$, ..., $\xi_n = x_n-a$, we have

$$\frac{d\xi_1}{da}=\frac{d\xi_2}{da}=\ldots=\frac{d\xi_n}{da}=-1.$$

Hence
$$\frac{\phi'(\xi_1)}{\phi(\xi_1)}+\frac{\phi'(\xi_2)}{\phi(\xi_2)}+\ldots+\frac{\phi'(\xi_n)}{\phi(\xi_n)}=0. \qquad (2)$$

Since, by hypothesis, a is the average of $x_1, x_2, \ldots, x_n$,

$$\xi_1+\xi_2+\ldots+\xi_n=0. \qquad (3)$$

Combining (2) and (3) we have

$$\left(\frac{\phi'(\xi_1)}{\phi(\xi_1)}-\lambda\xi_1\right)+\left(\frac{\phi'(\xi_2)}{\phi(\xi_2)}-\lambda\xi_2\right)+\ldots+\left(\frac{\phi'(\xi_n)}{\phi(\xi_n)}-\lambda\xi_n\right)=0,$$

where λ is any constant. Since $\xi_1, \xi_2, \ldots, \xi_n$ are subject only to the relation (3), it follows from this equation that

$$\frac{1}{\xi_1}\frac{\phi'(\xi_1)}{\phi(\xi_1)}=\frac{1}{\xi_2}\frac{\phi'(\xi_2)}{\phi(\xi_2)}=\ldots=\frac{1}{\xi_n}\frac{\phi'(\xi_n)}{\phi(\xi_n)}=\lambda.$$

These equations are all particular cases of the equation

$$\frac{\phi'(\xi)}{\phi(\xi)}=\lambda\xi,$$

the integral of which is $\phi(\xi) = Ae^{\lambda\xi^2/2}$. Apart from the constants λ and A, this determines the general form of the probability function which we have been led to by our assumptions.

It remains to determine whether we can choose λ so as to make P a maximum and not a minimum. Since P will be a maximum when $\log P$ is a maximum, P must satisfy the further condition that

$$\frac{d^2}{da^2}(\log P) < 0, \text{ for the specified value of } a.$$

Now this requires that

$$\sum_{r=1}^{n}\left\{\frac{\phi(\xi_r)\phi''(\xi_r)-\phi'(\xi_r)^2}{\phi(\xi_r)^2}\right\} < 0,$$

where
$$\phi(\xi) = Ae^{\lambda\xi^2/2}, \quad \phi'(\xi) = \lambda\xi\phi(\xi),$$
$$\phi''(\xi) = \lambda\xi\phi'(\xi)+\lambda\phi(\xi).$$

Thus the condition reduces to

$$\sum_{r=1}^{n}\lambda < 0,$$

so that λ must be negative. We then write

$$\phi(\xi) = Ae^{-h^2\xi^2}.$$

Since $\phi(\xi)$ is a probability function its total integral over the range of variation of ξ must be unity. Evidently our assumptions have led us to a function which does not vanish outside a finite range for ξ, but which admits the possibility of observations differing from the average by any number, however great. Clearly this result is a violation of the most elementary practice in observational work and is thus a measure of the extent to which assumptions (1) and (2) lead to hypothetical populations that are not consistent with practice. We shall discuss these limitations later; for the moment we use the fact that the range of the variable ξ must be taken as extending from $-\infty$ to $+\infty$. Hence we have

$$\int_{-\infty}^{\infty} Ae^{-h^2\xi^2}\,d\xi = 1,$$

from which it follows that $A = h/\sqrt{\pi}$ (p. 123).

Thus $\phi(\xi) = (h/\sqrt{\pi})e^{-h^2\xi^2}$, the Gaussian error law.

Alternative Derivation of the Gaussian Law

Another method of obtaining the Gaussian error law rests on assumptions of a different character. Let us seek a probability distribution in two dimensions which is a function of the radius vector only; that is, if x and y are the Cartesian coordinates, the required function is of the form $\phi(r)$, where $r^2 = x^2+y^2$.

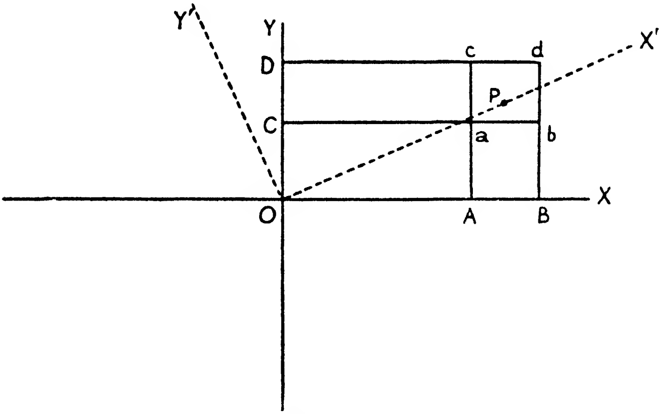


Fig. 21

Let P be a point (x, y) distant r from the origin O, and situated at the centre of a small square *abdc* of side α which is formed by drawing parallels to the axes through the points $ABCD$.

The probability that a point will lie in the annulus defined by two circles with centre at O and radii r, $r+dr$ is $\phi(r)\,dr$. Thus the probability that a point (x, y) will lie in the interval AB is $\phi(x)\alpha$; similarly, the probability that it will lie in the interval CD is $\phi(y)\alpha$.

We now assume that the probability that a point will lie inside the square *abdc* is the compound probability arising from these two independent events, i.e. $\phi(x)\phi(y)\alpha^2$. This result must remain unaltered if the axes OX, OY are rotated into the positions OX', OY'. If we construct a small square of side α, as in the previous case, we thus obtain

$$\phi(x)\phi(y)\alpha^2 = \phi(x')\phi(y')\alpha^2.$$

If the axes OX', OY' are so chosen that OX' passes through

$P(x,y)$, we have $x' = \sqrt{(x^2+y^2)}$, $y' = 0$. Thus the equation to be satisfied by the function is

$$\phi(x)\phi(y) = \phi\{\sqrt{(x^2+y^2)}\}\phi(0).$$

Assuming that ϕ is differentiable and differentiating first with respect to x and then with respect to y, we have

$$\phi'(x)\phi(y) = \frac{x}{\sqrt{(x^2+y^2)}}\phi'\{\sqrt{(x^2+y^2)}\}\phi(0),$$

$$\phi(x)\phi'(y) = \frac{y}{\sqrt{(x^2+y^2)}}\phi'\{\sqrt{(x^2+y^2)}\}\phi(0).$$

Hence $y\phi'(x)\phi(y) = x\phi(x)\phi'(y)$, or

$$\frac{\phi'(x)}{x\phi(x)} = \frac{\phi'(y)}{y\phi(y)}.$$

Since x and y are independent variables, this equality can hold only if both terms are constant; thus

$$\frac{\phi'(x)}{x\phi(x)} = A; \qquad \frac{\phi'(y)}{y\phi(y)} = A.$$

Hence $$\log\phi(x) = \frac{Ax^2}{2}+B,$$

or $$\phi(x) = Ce^{Dx^2},$$

where C and D are arbitrary constants. We have thus determined the nature of the function ϕ. We have still to insert the condition that the total area between the probability curve and the axis of X is unity.

Before doing so let us notice one consequence of our assumption that the probability of a point P falling inside the square *abdc* is equal to $\phi(x)\phi(y)\alpha^2$. It is clear that no probability function which was zero outside a circle of finite radius R could satisfy this condition, since there exist points lying outside this circle which have x- or y-coordinates of magnitude less than R; thus we should require the product of two finite quantities to be zero. It follows that our assumption cannot apply to a continuous function $\phi(r)$ which vanishes for values of $r > R$. In fact $\phi(x)$ must be finite for all finite values of x, as follows also from the result

$$\phi(x) = Ce^{Dx^2}.$$

By choosing D to be negative we can, however, make $\phi(x)$ decrease rapidly as x increases. We write

$$\phi(x) = Ce^{-h^2x^2},$$

where C and h are unspecified real numbers. We now apply the condition that the area between the probability curve and the x-axis is unity; since the range of x is $(-\infty, \infty)$ we thus obtain

$$\int_{-\infty}^{\infty} \phi(x)\,dx = 1.$$

Hence

$$1 = C\int_{-\infty}^{\infty} e^{-h^2x^2}\,dx = 2C\int_{0}^{\infty} e^{-h^2x^2}\,dx = \frac{2C}{h}\int_{0}^{\infty} e^{-z^2}\,dz,$$

where we have written $hx = z$.

It can be shown that

$$\int_{0}^{\infty} e^{-z^2}\,dz = \frac{\sqrt{\pi}}{2}.$$

Thus

$$\frac{C\sqrt{\pi}}{h} = 1, \quad \text{or} \quad C = \frac{h}{\sqrt{\pi}}.$$

Finally, therefore, we have

$$\phi(x) = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}.$$

Consider the expression

$$\int_{-\infty}^{\infty} \frac{h}{\sqrt{\pi}}e^{-h^2x^2}x^2\,dx = \frac{2}{h^2\sqrt{\pi}}\int_{0}^{\infty} e^{-z^2}z^2\,dz,$$

where $hx = z$. Now

$$\int_{0}^{\infty} e^{-z^2}z^2\,dz = \int_{0}^{\infty} ze^{-z^2}z\,dz = \left[-\frac{ze^{-z^2}}{2}\right]_0^{\infty} + \frac{1}{2}\int_{0}^{\infty} e^{-z^2}\,dz = \frac{\sqrt{\pi}}{4}.$$

Thus

$$\int_{-\infty}^{\infty} x^2\phi(x)\,dx = \frac{1}{2h^2}.$$

We have seen that if we have a set of observations such that x is the difference of an observation from the average, and if $\phi(x)$ is the frequency with which x occurs in the set, then $\int_{-\infty}^{\infty} x^2\phi(x)\,dx$ is an approximation to σ^2, the square of the standard deviation. It follows that

$$\sigma^2 = \frac{1}{2h^2}, \quad \text{or} \quad h = \frac{1}{\sigma\sqrt{2}}, \text{ approximately.}$$

The probability function then takes the form

$$\phi(x) = \frac{1}{\sigma\sqrt{(2\pi)}}e^{-x^2/2\sigma^2},$$

where σ is the standard deviation of the hypothetical population.

The Error Function

The Gaussian probability curve, given by $y = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}$, is shown roughly in the accompanying diagram.

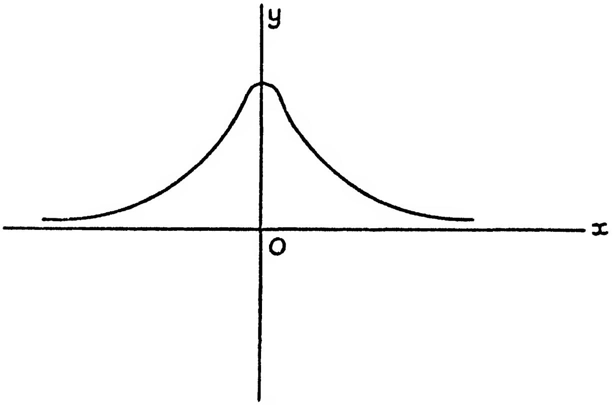


Fig. 22

It has a maximum at $x = 0$, of amount $h/\sqrt{\pi}$, and points of inflexion, found by writing $d^2y/dx^2 = 0$, at the points $x = \pm 1/h\sqrt{2}$. We have already shown above that $1/h\sqrt{2} = \sigma'$, a quantity which, for the Gaussian law, corresponds to the standard deviation σ for a finite set of observations.

It is clear that the greater the value of h, the more closely does the curve lie to the x-axis; thus it is suggested that the constant h is associated with the precision of any set of data which might conform to the Gaussian law.

The probability that a variable will have a deviation between x and $x+dx$ is $\frac{h}{\sqrt{\pi}}e^{-h^2x^2}\,dx$; thus the probability that a deviation will lie in the range (a,b) is $\frac{h}{\sqrt{\pi}}\int\limits_a^b e^{-h^2x^2}\,dx$. The probability that a variable will have a deviation between $-1/h\sqrt{2}$ and $1/h\sqrt{2}$, the positions of the inflexions, is

$$p = \frac{h}{\sqrt{\pi}}\int\limits_{-\frac{1}{h\sqrt{2}}}^{\frac{1}{h\sqrt{2}}} e^{-h^2x^2}\,dx = \frac{2}{\sqrt{\pi}}\int\limits_0^{\frac{1}{\sqrt{2}}} e^{-t^2}\,dt, \text{ where } t = hx.$$

Evaluation of the Error Function

Because of its importance for probabilities whose frequencies are given by the Gaussian law, the error function $\frac{2}{\sqrt{\pi}}\int\limits_0^x e^{-x^2}\,dx$ has been studied in detail and tabulated (see Appendix). Various methods have been adopted for this purpose. We notice, in the first place, that a particular value of the function is known, for $\int\limits_0^\infty e^{-x^2}\,dx = \frac{1}{2}\sqrt{\pi}$. Thus, as in Chapter V, if we write

$$\text{Erf}(x) = \frac{2}{\sqrt{\pi}}\int\limits_0^x e^{-x^2}\,dx,$$

then $\quad \text{Erf}(\infty) = 1, \quad \text{and} \quad \text{Erf}(0) = 0.$

When x is small we may approximate to the value of Erf(x) as follows.

We have

$$\int\limits_0^x e^{-x^2}\,dx = \int\limits_0^x \left[1-x^2+\frac{x^4}{2!}-\frac{x^6}{3!}+\ldots\right]dx$$

$$= x-\frac{x^3}{3}+\frac{x^5}{5.2!}-\frac{x^7}{7.3!}+\ldots.$$

Since the series is an alternating one, the sum to two successive terms gives an upper and lower limit to its sum. Thus, if we reject the terms beyond x^7, the result will be deficient by an amount less than $\dfrac{x^9}{9.4!}$. If this is to be less than unity in the fourth decimal place, we require

$$\frac{x^9}{9.4!} < 10^{-4}, \quad \text{or} \quad x < 2.10^{-2}, \text{ approximately.}$$

For large values of x we proceed differently. Integrating by parts, we have

$$\int\limits_x^\infty e^{-x^2}\,dx = \int\limits_x^\infty \frac{1}{x}xe^{-x^2}\,dx = \frac{1}{2x}e^{-x^2} - \frac{1}{2}\int\limits_x^\infty \frac{1}{x^2}e^{-x^2}\,dx$$

$$= \frac{1}{2x}e^{-x^2} - \frac{1}{2^2x^3}e^{-x^2} + \frac{1.3}{2^2}\int\limits_x^\infty \frac{1}{x^4}e^{-x^2}\,dx.$$

Continuing this process, we obtain

$$\int\limits_x^\infty e^{-x^2}\,dx = \frac{e^{-x^2}}{2x}\left\{1 - \frac{1}{2x^2} + \frac{1.3}{(2x^2)^2} - \frac{1.3.5}{(2x^2)^3} + \ldots\right\}.$$

Since the function e^{-x^2} is decreasing in the range (x, ∞) it is clear that the error involved in stopping at the fourth term is less than $e^{-x^2}\int\limits_x^\infty \dfrac{1.3.5.7}{2^4x^8}\,dx$, numerically, i.e. less than $e^{-x^2}\dfrac{1.3.5}{2^4x^7}$, the last term retained. A similar result is obtained at any stage of the expansion.

The Probable Error

We define the probable error for a Gaussian distribution in analogy with that of a finite set of observations by stating that it is a deviation the probability of whose occurrence is $\frac{1}{2}$. Thus, if r is the probable error, then

$$\frac{h}{\sqrt{\pi}}\int\limits_{-r}^{r} e^{-h^2x^2}\,dx = \tfrac{1}{2}.$$

As in previous examples, we express this integral in terms of Erf x. Writing $h = 1/\sigma\sqrt{2}$, we require the value of r which makes

$$\frac{2}{\sqrt{\pi}}\int\limits_0^{r/\sigma\sqrt{2}} e^{-z^2}\,dz = \tfrac{1}{2},$$

where $hx = z$. Thus $\text{Erf}(r/\sigma\sqrt{2}) = 0{\cdot}5$.

From the table we find that

$$r/\sigma\surd 2 = 0{\cdot}477,$$

and $$r = 0{\cdot}6745\sigma.$$

As we have seen, the probable error gives the upper and lower limits for the deviation of a variable such that the probability of the deviation lying within those limits is equal to the probability for which it lies outside. Or we may say: the odds in favour of the deviation lying within the range $\pm r$ are 1 : 1. We may inquire what are the odds in favour of the deviation lying in the ranges $\pm 2r$, $\pm 3r$....

Thus, the probability that the deviation will lie in the range $\pm 2r$ is

$$\frac{2}{\surd\pi}\int\limits_0^{2r/\sigma\surd 2} e^{-z^2}\,dz, \quad \text{or} \quad \mathrm{Erf}(2r/\sigma\surd 2).$$

Since $r/\sigma\surd 2 = 0{\cdot}477$, the probability is $\mathrm{Erf}(0{\cdot}954) = 0{\cdot}83$, from the tables.

Hence the odds in favour of this range are

$$0{\cdot}83 : 1-0{\cdot}83 = 9 : 2, \text{ approximately.}$$

Ex. 1. Show that the approximate odds in favour of a deviation lying in the range

$\pm 3r$ are 21 : 1;
$\pm 4r$ are 142 : 1;
$\pm 5r$ are 1,310 : 1;
$\pm 6r$ are 19,200 : 1;
$\pm 7r$ are 420,000 : 1;
$\pm 8r$ are $17\times 10^6 : 1$;
$\pm 9r$ are $10^8 : 1$.

Ex. 2. What is the probability that a deviation will lie in the range $\pm\sigma$?

The required probability is $\dfrac{2}{\sigma\surd\pi}\displaystyle\int\limits_0^{\sigma} \exp(-x^2/2\sigma^2)\,dx = \mathrm{Erf}\,\frac{1}{\surd 2}.$

From the table, $\mathrm{Erf}\,(1/\surd 2) = 0{\cdot}682$.

Thus the odds in favour are $0{\cdot}682 : 1-0{\cdot}682 = 17 : 8$, approximately.

Show that, for a range $\pm 2\sigma$, the probability is 0·954,
$\pm 3\sigma$, the probability is 0·997,
$\pm 4\sigma$, the probability is 0·99994.

Applications of the Normal Law

If the probabilities of the occurrence of x and y, two numbers in the range $-\infty < (x,y) < +\infty$ are respectively

$$\frac{h}{\sqrt{\pi}}\exp(-h^2x^2) \text{ and } \frac{k}{\sqrt{\pi}}\exp(-k^2y^2),$$

x and y being chosen independently, we require the probability that $f(x,y)$ lies in the range

$$\mu \leqslant f(x,y) \leqslant \mu+\delta\mu.$$

The compound probability P is clearly

$$\frac{h}{\sqrt{\pi}}\exp(-h^2x^2).\frac{k}{\sqrt{\pi}}\exp(-k^2y^2)\,dxdy,$$

integrated over the range of x and y specified by the above inequality.

Consider as an illustration the case

$$f(x,y) = x+y,$$

i.e. $$\mu \leqslant x+y \leqslant \mu+\delta\mu.$$

Then $$P = \frac{hk}{\pi}\int\limits_{-\infty}^{\infty} \exp(-h^2x^2)\,dx \int\limits_{\mu-x}^{\mu+\delta\mu-x} \exp(-k^2y^2)\,dy.$$

Now by the mean ordinate rule for integration we may write

$$\int\limits_{\mu-x}^{\mu+\delta\mu-x} \exp(-k^2y^2)\,dy$$
$$= \tfrac{1}{2}[\exp\{-k^2(\mu+\delta\mu-x)^2\}+\exp\{-k^2(\mu-x)^2\}]\,\delta\mu$$
$$= \exp\{-k^2(\mu-x)^2\}\,\delta\mu.$$

Hence $$P = \frac{hk}{\pi}\int\limits_{-\infty}^{\infty} \exp\{-h^2x^2-k^2(\mu-x)^2\}\,dx\delta\mu.$$

Now
$$h^2x^2+k^2(\mu-x)^2 = (h^2+k^2)\left(x-\frac{k^2\mu}{h^2+k^2}\right)^2+\frac{h^2k^2}{h^2+k^2}\mu^2.$$

Hence

$$P = \frac{hk\,\delta\mu}{\pi}\exp\left(-\frac{h^2k^2}{h^2+k^2}\mu^2\right)\int\limits_{-\infty}^{\infty}\exp\left\{-(h^2+k^2)\left(x-\frac{k^2\mu}{h^2+k^2}\right)^2\right\}dx$$

$$= \frac{hk}{\sqrt{(h^2+k^2)}}\frac{\delta\mu}{\sqrt{\pi}}\exp\left(-\frac{h^2k^2}{h^2+k^2}\mu^2\right)$$

or $$P = \frac{l}{\sqrt{\pi}}\delta\mu\exp(-l^2\mu^2),$$

where $$\frac{1}{l^2} = \frac{1}{h^2}+\frac{1}{k^2}.$$

Following precisely the same line of development it is easily verified that if $f(x,y) = ax+by$ then the probability that $ax+by$ is chosen in the range $(\mu, \mu+\delta\mu)$ is

$$P = \frac{L}{\sqrt{\pi}}\exp(-L^2\mu^2)\,\delta\mu,$$

where $$\frac{1}{L^2} = \frac{a^2}{h^2}+\frac{b^2}{k^2}.$$

Once again this is easily generalized to the following proposition:

If $x_1, x_2,..., x_n$ be a set of n independently chosen numbers in the range $(\infty, -\infty)$, and if the probabilities with which $x_1, x_2, x_3,...$ are chosen are

$$\frac{h_1}{\sqrt{\pi}}\exp(-h_1^2x_1^2),\quad \frac{h_2}{\sqrt{\pi}}\exp(-h_2^2x_2^2),\quad ...,\quad \frac{h_n}{\sqrt{\pi}}\exp(-h_n^2x_n^2),$$

then the probability that

$$a_1x_1+a_2x_2+...+a_nx_n$$

shall lie in the range $(\mu, \mu+\delta\mu)$ is

$$\frac{l\,\delta\mu}{\sqrt{\pi}}\exp(-l^2\mu^2),$$

where $$\frac{1}{l^2} = \frac{a_1^2}{h_1^2}+\frac{a_2^2}{h_2^2}+...+\frac{a_n^2}{h_n^2}.$$

Accuracy of the Arithmetic Mean

Let $$a_1 = a_2 = ... = a_n = 1/n,$$

then the probability of $(x_1+x_2+...+x_n)/n$ lying in the range $(\mu, \mu+\delta\mu)$ is

$$\frac{l\,\delta\mu}{\sqrt{\pi}}\exp(-l^2\mu^2),$$

where $$\frac{1}{l^2} = \frac{1}{n^2}\left[\frac{1}{h_1^2}+\frac{1}{h_2^2}+\ldots+\frac{1}{h_n^2}\right].$$

If all the quantities h_r have equal values, then

$$\frac{1}{l^2} = \frac{n}{h^2}\frac{1}{n^2} = \frac{1}{nh^2} \text{ or } l = h\sqrt{n}.$$

Accordingly the required probability is

$$\frac{h\sqrt{n}}{\sqrt{\pi}}\exp(-nh^2\mu^2)\,\delta\mu.$$

The fact that all the quantities h_r are equal implies that all the measures $x_1,\ldots, x_n$ are equally precise, i.e. they each belong to groups having the same standard deviation

$$\sigma = \frac{1}{h\sqrt{2}}.$$

Now from the composite law of error of the arithmetic mean,

$$P = \frac{l\,\delta\mu}{\sqrt{\pi}}\exp(-l^2\mu^2),$$

and therefore the standard deviation for the arithmetic mean is

$$S = \frac{1}{l\sqrt{2}} = \frac{1}{h\sqrt{(2n)}} = \frac{\sigma}{\sqrt{n}}.$$

Thus *the accuracy of the arithmetic mean of n observations is $\sqrt{n}$ times that of a single observation of the system, if all are equally good and if the deviations of the observations and of the means satisfy the Gaussian law.*

Ex. Consider the probability that $(x^2+y^2)^{\frac{1}{2}}$ lies between μ and $\mu+\delta\mu$ when x and y are selected in the range $(-\infty, \infty)$ according to the Gaussian law, with equal precision constants. Here

$$P = \iint \frac{h^2}{\pi}\exp\{-h^2(x^2+y^2)\}\,dxdy,$$

the integral extending over the region defined by

$$\mu+\delta\mu \geqslant (x^2+y^2)^{\frac{1}{2}} \geqslant \mu.$$

Thus

$$P = \frac{h^2}{\pi}\int\limits_0^{2\pi} d\theta \int\limits_{\mu}^{\mu+\delta\mu} r\exp(-h^2r^2)\,dr = [-\exp(-h^2r^2)]_{\mu}^{\mu+\delta\mu}$$
$$= \exp(-h^2\mu^2)-\exp\{-h^2(\mu+\delta\mu)^2\} = 2h^2\mu\,\delta\mu\exp(-h^2\mu^2).$$

Hence, also, the probability of $\sqrt{(x^2+y^2)}$ lying between μ_1 and μ_2 is

$$\int\limits_{\mu_1}^{\mu_2}\exp(-h^2\mu^2)2h^2\mu\,d\mu = \exp(-h^2\mu_1^2)-\exp(-h^2\mu_2^2).$$

The Random Walk

On p. 81 we dealt with the problem of the Random Walk in two dimensions where the length of each walk was specified but the direction undetermined, all directions having an equal probability. We turn now to an examination of the complementary problem, in which the directions are specified but the distances traversed in each direction are undetermined except that they are each drawn, as it were, from stocks distributed about the mean, according to the Gaussian law. We consider, therefore, the simple case of two component translations x and y at right angles.

An individual walks a distance x from a point O, then turning at right angles walks a distance y. If the probability that x lies between x and $x+\delta x$ is $\frac{h}{\sqrt{\pi}}\exp(-h^2x^2)\,\delta x$ and that y lies between y and $y+\delta y$ is $\frac{k}{\sqrt{\pi}}\exp(-k^2y^2)\,\delta y$, it is required to determine the probability that the individual is finally to be found at a distance between μ and $\mu+\delta\mu$ from O.

Since x and y are not selected with equal precision, but according to the laws $(h/\sqrt{\pi})\exp(-h^2x^2)$ and $(k/\sqrt{\pi})\exp(-k^2y^2)$, then

$$P = \frac{hk}{\pi}\int\limits_{-\mu}^{\mu}\exp(-h^2x^2)\,dx \times \int\limits_{\sqrt{\{\mu^2-x^2\}}}^{\sqrt{\{(\mu+\delta\mu)^2-x^2\}}}\exp(-k^2y^2)\,dy.$$

The limits of integration are determined from the fact that

$$\mu \leqslant \sqrt{(x^2+y^2)} \leqslant \mu+\delta\mu,$$

i.e.
$$\sqrt{(\mu^2-x^2)} \leqslant y \leqslant \sqrt{[(\mu+\delta\mu)^2-x^2]}.$$

Now
$$\sqrt{[(\mu+\delta\mu)^2-x^2]} = \sqrt{(\mu^2-x^2+2\mu\,\delta\mu+\delta\mu^2)}$$
$$= \sqrt{(\mu^2-x^2)}\left[1+\frac{\mu\,\delta\mu}{\mu^2-x^2}+\ldots\right]$$
$$= \sqrt{(\mu^2-x^2)}+\frac{\mu\,\delta\mu}{\sqrt{(\mu^2-x^2)}},$$

on retaining terms of the first order in $\delta\mu$.

Thus the integral on the right becomes, since the two limits are close together,

$$[\exp\{-k^2((\mu+\delta\mu)^2-x^2)\}+\exp\{-k^2(\mu^2-x^2)\}]\frac{\mu\,\delta\mu}{2\sqrt{(\mu^2-x^2)}}$$
$$= \frac{\mu\,\delta\mu}{\sqrt{(\mu^2-x^2)}}\exp\{-k^2(\mu^2-x^2)\}.$$

Accordingly,

$$P = \frac{\mu hk}{\pi}\delta\mu \int\limits_{-\mu}^{\mu} \exp(-h^2x^2)\frac{\exp\{-k^2(\mu^2-x^2)\}}{\sqrt{(\mu^2-x^2)}}\,dx.$$

Let $x = \mu\sin\theta$; then

$$P = \frac{\mu hk\,\delta\mu}{\pi} \int\limits_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \exp\{-(h^2\sin^2\theta+k^2\cos^2\theta)\mu^2\}\,d\theta$$

$$= \frac{\mu hk}{\pi}\exp\left(-\frac{h^2+k^2}{2}\mu^2\right)\delta\mu \int\limits_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \exp\left(\mu^2\frac{h^2-k^2}{2}\cos 2\theta\right)d\theta.$$

Now if $\lambda^2 = \frac{1}{2}\mu^2(k^2-h^2)$, then

$$\int\limits_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \exp(-\lambda\cos 2\theta)\,d\theta = \frac{1}{2}\int\limits_{-\pi}^{\pi} \exp(-\lambda\cos\phi)\,d\phi, \quad \text{where } \phi = 2\theta,$$

$$= \frac{1}{2}\int\limits_{-\pi}^{\pi}\left(1-\lambda\cos\phi+\frac{\lambda^2}{2!}\cos^2\phi-\ldots\right)d\phi$$

$$= 2\int\limits_{0}^{\frac{1}{2}\pi}\left(1+\frac{\lambda^2}{2!}\cos^2\phi+\frac{\lambda^4}{4!}\cos^4\phi+\ldots\right)d\phi$$

$$= \pi\left[1+\frac{\lambda^2}{2^2}+\frac{1}{(2!)^2}\frac{\lambda^4}{2^4}+\frac{1}{(3!)^2}\frac{\lambda^6}{2^6}+\ldots\right]$$

$$= \pi J_0(\lambda i) = \pi J_0[\mu\sqrt{(h^2-k^2)}/\sqrt{2}].$$

Thus, finally, since μ is regarded as positive,

$$P = 2\mu hk\exp\{-\tfrac{1}{2}(h^2+k^2)\mu^2\}J_0[\mu\sqrt{(h^2-k^2)}/\sqrt{2}]\,\delta\mu.$$

We note that if $h = k$, then $\lambda = 0$, and

$$P = 2\mu h^2\exp(-\mu^2h^2)\,\delta\mu.$$

This problem finds an interesting application in the determination of persistent periodicities in observations. (See J. Bartels, *Terrestrial Magnetism*, etc., vol. 40, no. 1, 1935.)

Ex. 1. Particles are distributed in a plane X, Y in such a manner that their x- and y-coordinates belong to Gaussian sets of standard deviation σ.

Show that the probability that the distance from O of any one of them lies (*a*) between 0 and σ is $(\sqrt{e}-1)/\sqrt{e}$, (*b*) between $\alpha\sigma$ and $\beta\sigma$ is $e^{-\beta^2/2}-e^{-\alpha^2/2}$, where $\alpha < \beta$.

Ex. 2. If in the foregoing example the 'probable distance R' from O be defined as that for which it is equally probable that the particle will lie within it as without, show that

$$R^2 = \sigma^2\log_e 4.$$

Show further that the region of greatest density of particles is in the neighbourhood of $r = \sigma$.

The Gaussian Law and Experiment

At this stage it is worth while reviewing again the position of the Gaussian law with regard to experimental observation. The law has been derived by us on assumptions which cannot be held to apply rigorously in practice (p. 120); moreover, like the Bernoulli law, the Gaussian law indicates what frequency curve will be found on these assumptions when all possible arrangements of the elements considered have been included. Now it is always possible to assume that any frequency curve obtained in practice represents a sample of a super-population; it can be regarded as a selected and not an exhaustive collection of the possible arrangements. This has to be borne in mind if we are not to apply the Gaussian law uncritically. But there is another and, in a sense, more fundamental objection: it may not be true—and there is no reason to suppose it even approximately true—that all the arrangements of data which might be chosen from the population necessarily show that the hypothetical population conforms to a Gaussian distribution.

When a set of data does not so conform, one is tempted to assert that this circumstance arises from the fact that the data represent only a sample; but it may be that the original population is *not* Gaussian. The position is clearly seen from the investigation on p. 156; it is there shown that if a population has its frequency expressible as a function $v(t)$ of a variable t representing some characteristic, and if in sampling the population at what is presumed to be a value t, we draw in sets of data in the neighbourhood of t, with a probability of choice $p(x)$ at $t+x$, then the final sampling distribution is given by

$$u(t) = \int v(t+x)p(x)\,dx,$$

the integral extending over the range of the sampling. It is clear that the form of a sample depends on the conjunction of the distribution in the population and the law of choice over the range specified. As we shall prove, when the population and the law of choice are Gaussian, so is the sample; but if either the population or the law of choice be not Gaussian, the sample is not Gaussian. It follows that to apply conclusions drawn from a Gaussian distribution to the interpretation of *any* group of samples may involve us in serious error.

Here again we must not attempt to escape from this impasse by asserting that, in the last resort, the Gaussian law gives an idealized distribution by which to interpret any given set of data. There is no escaping the plain issue that every such interpretation must stand side by side with the *assumption* that the original population is Gaussian.

The Significance of Deviations

In connexion with the above remarks we may consider generally the problem of significance as it arises in statistical theory. Broadly speaking, we may say that the significance of a statistical constant is usually estimated by comparing it with the corresponding constant which would be found under so-called 'conditions of randomness'; that is, by calculating the probability that a constant of this magnitude would be found under conditions in which all possible arrangements could occur. Thus, let us suppose that certain data are presumed to be measurements carried out under the same physical conditions on the same object, and that the deviations from the average have been found and the standard deviation σ calculated for these observations. So far we have made no assumption regarding the nature of any distribution law to which the measurements are presumed to conform. Now suppose that one of them in particular appears to differ very considerably from the others, showing a deviation 4σ, say—the deviation having been found by *including* this observation. A good experimenter may justifiably have his suspicions aroused as to the accuracy of the observation: how is he to decide whether it should be included or not? If he is a sensible experimenter he will know whether any suspicion arising in the course of his work attaches itself to this particular observation—no statistician could possibly tell him that—and in so far as he relegates his judgement on this matter to the statistician he is surrendering his function as an experimenter. All that the statistician can tell him is how far the set of measurements are consistent with some assumed law of distribution, for there is no meaning in the bald statement, 'the numbers are consistent among themselves'. Thus, what the statistician does is to seek the probability that a deviation from the average as large as 4σ will be found from the same number of data drawn 'at random' from an original population, the structure of which he proceeds to specify.

The experimenter has no such knowledge of this structure: one of the purposes of his experiment is to find it. As we have seen, the odds against a deviation of 4σ, on the Gaussian law, are about $10^5 : 6$ or $10^4 : 1$, approximately. And if the experimenter is overwhelmed by this fact he accepts without further question the significance of the odds. Thus, the significance of the observation is referred by this process to the significance of a probability arising from an assumed population and, accordingly, the experimenter may decide to reject this observation. The statistician does not in fact do precisely this; he states that when the odds are, say, 25 : 1 against, he will advise rejection. The justification of this judgement is stated to be based on experience; but if it is, it can only be the experience of the experimenter reinterpreted by the statistician.

3. Other forms of hypothetical populations

In general we may say that when from a set of data, restricted in extent, a frequency or probability curve is constructed and its equation expressed by a mathematical formula, we have thereby invented a hypothetical population of which our data may be regarded as samples. There is clearly a considerable latitude in specifying this formula; the mathematician knows that through a finite number of points will pass an infinity of curves, so that other conditions describing the nature of the formula to be used must be given before we can assert that the final result represents *the* hypothetical population which satisfies our requirements. This problem of constructing the hypothetical population is simply a restatement of the abovementioned problem of determining the original population when the data and the method by which they were selected are known; if, of course, no method of selection is specified, then all sorts of formulae can be found. A given type of formula for the population implies some kind of selective process, even if it is not explicitly stated.

In the notation of the previous section, the sample $u(t)$ of a population $v(t)$ defined in the range $(-\infty, \infty)$ is given by the equation

$$u(t) = \int_{-\infty}^{\infty} v(t+x)p(x)\,dx. \tag{1}$$

Now assume that $v(t)$ can be expanded in a Taylor series

$$v(t+x) = v(t)+xv'(t)+\frac{x^2}{2!}v''(t)+\ldots. \tag{2}$$

If we introduce the constants† m_1, m_2,..., m_r defined by the relations

$$m_1 = \int_{-\infty}^{\infty} xp(x)\,dx, \qquad m_2 = \frac{1}{2!}\int_{-\infty}^{\infty} x^2p(x)\,dx, \qquad \ldots,$$

$$m_r = \frac{1}{r!}\int_{-\infty}^{\infty} x^rp(x)\,dx, \tag{3}$$

we may write (1) in the form

$$u(t) = v(t)+m_1v'(t)+m_2v''(t)+\ldots. \tag{4}$$

That is, the sample can be expressed in terms of the probability function for the hypothetical population, its derivatives, and the moment coefficients of the probability function of selection.

We observe that if the function $p(x)$ is a symmetrical (i.e. an even) function, then the coefficient m_r is zero for all odd values of r. In this case the sample $u(t)$ will be expressible in terms of $v(t)$ and its even derivatives.

Ex. If $p(x) = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}$, show that

$$m_{2r} = \frac{1}{4rh^2}m_{2r-2} = \frac{1}{(2h)^{2r}r!}.$$

It is a simple matter to invert equation (4), supposing that the operation of inversion is permissible. For we have, by successive approximation,

$$v(t) = u(t)-m_1v'(t)-m_2v''(t)+\ldots,$$

or

$$v(t) = u(t)-m_1u'(t)-m_2\{u''(t)-m_1u'''(t)\}+\ldots. \tag{5}$$

† These are numerical multiples of the 'moments' of $p(x)$ as usually defined.

This formula expresses $v(t)$ in terms of the simple function $u(t)$ and its successive derivatives.

Ex. Show that, if $p(x) = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}$,

$$v(t) = u(t) - \frac{1}{4h^2}u''(t) + \frac{1}{32h^4}u^{\text{iv}}(t) + \dots.$$

If the hypothetical population is itself Gaussian, i.e. if $v(t)$ is of the form $\frac{h}{\sqrt{\pi}}e^{-h^2t^2}$, then irrespective of the method of selection, it follows from the foregoing that we should be able to expand a given sample function in a series of terms consisting of numerical multiples of $e^{-h^2x^2}$ and its successive derivatives. We may invert this process; in fact formula (1) shows that if a sample of a continuous variable is assumed to be Gaussian, then the hypothetical population can be expressed as a series of linear combinations of $e^{-h^2x^2}$ and its derivatives. In both cases the coefficients in the expansion are definite numerical multiples of the moment coefficients of the probability distribution used in the process of selecting the sample from the hypothetical population. It remains, therefore, to examine the procedure to be followed in order to expand a given function in the manner described.

The Hermite Polynomials

Consider the function $y = e^{-\frac{1}{2}x^2}$ (in which, for simplicity, we have written $h = 1/\sqrt{2}$ and omitted the factor $1/\sqrt{(2\pi)}$). The first derivative of y with regard to x is

$$\frac{dy}{dx} = -xe^{-\frac{1}{2}x^2}.$$

The second derivative is

$$\frac{d^2y}{dx^2} = -\frac{d}{dx}(xe^{-\frac{1}{2}x^2}) = (x^2-1)e^{-\frac{1}{2}x^2};$$

and in general the nth derivative is

$$\frac{d^ny}{dx^n} = (-1)^n\left\{x^n - \frac{n(n-1)}{2}x^{n-2} + \right.$$
$$\left. + \frac{n(n-1)(n-2)(n-3)}{2.4}x^{n-4} - \dots\right\}e^{-\frac{1}{2}x^2}. \qquad (6)$$

The expression

$$x^n-\frac{n(n-1)}{2}x^{n-2}+\frac{n(n-1)(n-2)(n-3)}{2.4}x^{n-4}-\dots,$$

which occurs in (6) is called the Hermite polynomial of order n, and is denoted by $H_n(x)$. It is easily shown that $H_n(x)$ satisfies the differential equation

$$\frac{d^2H_n(x)}{dx^2}-x\frac{dH_n(x)}{dx}+nH_n(x)=0, \tag{7}$$

and the recurrence relation

$$H_{n+1}(x)-xH_n(x)+nH_{n-1}(x)=0. \tag{8}$$

For since $$y=e^{-\frac{1}{2}x^2},$$

$$\frac{dy}{dx}=-xe^{-\frac{1}{2}x^2}=-xy.$$

Differentiating this result n times, we have

$$\frac{d^{n+1}y}{dx^{n+1}}+x\frac{d^ny}{dx^n}+n\frac{d^{n-1}y}{dx^{n-1}}=0,$$

which, since $\frac{d^ny}{dx^n}=(-1)^nH_n(x)y$, is equivalent to (8).

We have also

$$\frac{d^{n+2}y}{dx^{n+2}}+x\frac{d^{n+1}y}{dx^{n+1}}+(n+1)\frac{d^ny}{dx^n}=0,$$

or $$\frac{d^2}{dx^2}\{H_n(x)y\}+x\frac{d}{dx}\{H_n(x)y\}+(n+1)H_n(x)y=0.$$

Hence

$$\frac{d^2H_n(x)}{dx^2}y+\frac{2dH_n(x)}{dx}\frac{dy}{dx}+H_n(x)\frac{d^2y}{dx^2}+x\frac{dH_n(x)}{dx}y+$$
$$+xH_n(x)\frac{dy}{dx}+(n+1)H_n(x)y=0.$$

If we insert the values of dy/dx and d^2y/dx^2 and divide by y, this becomes

$$\frac{d^2H_n(x)}{dx^2}-2x\frac{dH_n(x)}{dx}+(x^2-1)H_n(x)+x\frac{dH_n(x)}{dx}-$$
$$-x^2H_n(x)+(n+1)H_n(x)=0,$$

which reduces to (7).

By means of (8) we can compute $H_n(x)$ for successive values of n, since $H_0(x)$ and $H_1(x)$ are both known.

It follows from the expression for d^ny/dx^n that the curves

$$y = \frac{d^{2r}}{dx^{2r}}(e^{-\frac{1}{2}x^2})$$

are all symmetrical, while the curves

$$y = \frac{d^{2r+1}}{dx^{2r+1}}(e^{-\frac{1}{2}x^2})$$

are skew (see Figs. 23, 24).

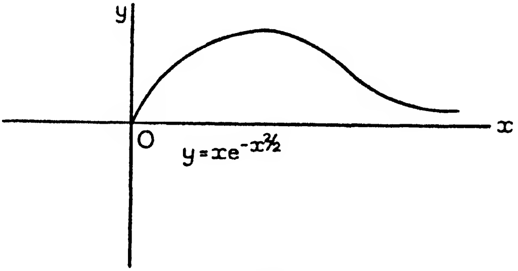


FIG. 23

It is clear that $y = xe^{-\frac{1}{2}x^2}$ can represent a probability curve, since

$$\int_0^\infty xe^{-\frac{1}{2}x^2}\,dx = \left[-e^{-\frac{1}{2}x^2}\right]_0^\infty = 1.$$

The mode or maximum value of y occurs when $x = 1$ and is thus $e^{-\frac{1}{2}}$.

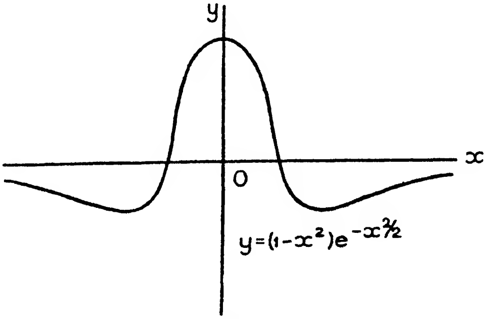


FIG. 24

The importance of the Hermite polynomials, from our point of view, lies in the fact that any given frequency function $f(x)$

which satisfies certain very general conditions† may be expanded in a series of the form

$$f(x) = a_0 e^{-\frac{1}{2}x^2} + a_1 e^{-\frac{1}{2}x^2} H_1(x) + a_2 e^{-\frac{1}{2}x^2} H_2(x) + \ldots \tag{9}$$

where a_0, a_1,... are constants.

To obtain the coefficient a_n, multiply both sides of this identity by $H_n(x)$. Integrating, we have

$$\int\limits_{-\infty}^{\infty} f(x) H_n(x)\, dx = \sum_{i=0}^{\infty} a_i \int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_i(x) H_n(x)\, dx. \tag{10}$$

Now

$$\int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m(x) H_n(x)\, dx = \left[(-1)^n H_m(x) \frac{d^{n-1}}{dx^{n-1}} (e^{-\frac{1}{2}x^2})\right]_{-\infty}^{\infty} - $$
$$- \int\limits_{-\infty}^{\infty} (-1)^n H_m'(x) \frac{d^{n-1}}{dx^{n-1}} (e^{-\frac{1}{2}x^2})\, dx,$$

on integration by parts,

$$= \int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m'(x) H_{n-1}(x)\, dx,$$

in virtue of the fact that the integrated part vanishes at both limits. Proceeding thus, we obtain, if $n > m$,

$$\int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m(x) H_n(x)\, dx = \int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m^{(m)}(x) H_{n-m}(x)\, dx. \tag{11}$$

Now, if $n = m$, we have

$$H_n^{(n)}(x) = n! \quad \text{and} \quad H_0(x) = 1.$$

Thus the integral reduces to $n! \int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = n! \sqrt{(2\pi)}$.

If, instead, $n > m$, integration of (11) gives us

$$\int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m(x) H_n(x)\, dx = \int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_m^{(m+1)}(x) H_{n-m-1}(x)\, dx. \tag{12}$$

But since $H_m(x)$ is a polynomial of degree m, $H_m^{(m+1)}(x)$ is zero; thus the left-hand side of (12) is also zero.

Returning now to (10), we see that if $i \neq n$, the coefficient

† It is sufficient that $f(x)$, $f'(x)$, and $f''(x)$ should be finite and continuous in $(-\infty, \infty)$ and that $f(x)$ and its derivatives should vanish at $x = \pm\infty$.

of a_i vanishes, by (12), while if $i = n$, it is equal to $n!\sqrt{(2\pi)}$. Hence we obtain from (10),

$$\int\limits_{-\infty}^{\infty} f(x)H_n(x)\,dx = a_n\,n!\sqrt{(2\pi)},$$

giving
$$a_n = \frac{1}{n!\sqrt{(2\pi)}}\int\limits_{-\infty}^{\infty} f(x)H_n(x)\,dx.$$

It will be observed that the method of determining a_n follows closely that of obtaining the coefficients in a Fourier expansion.

Thus, when the sample is expressed as a series of derivatives of $e^{-\frac{1}{2}x^2}$, the hypothetical population will itself be expressible in this form. The cases we have dealt with above are the comparatively simple ones in which one function or the other is Gaussian.

Standard Deviation for Bernoullian Populations

In this case the standard deviation σ from the average is given by (p. 62)

$$\sigma^2 = \sum_{r=0}^{n} {}^nC_r\,p^rq^{n-r}(np-r)^2$$
$$= n^2p^2\sum {}^nC_r\,p^rq^{n-r}-2np\sum {}^nC_r\,p^rq^{n-r}r+\sum {}^nC_r\,p^rq^{n-r}r^2. \tag{1}$$

From the identity

$$(p+q)^n = \sum {}^nC_r\,p^rq^{n-r},$$

we find as on p. 63 that

$$np(p+q)^{n-1} = \sum {}^nC_r\,p^rq^{n-r}r, \tag{2}$$

giving
$$np = \sum {}^nC_r\,p^rq^{n-r}r.$$

Differentiating (2) with respect to p,

$$n(p+q)^{n-1}+n(n-1)p(p+q)^{n-2} = \sum {}^nC_r\,p^{r-1}q^{n-r}r^2. \tag{3}$$

Hence, substituting from (2) and (3) in (1) we obtain

$$\sigma^2 = n^2p^2-2n^2p^2+np+n(n-1)p^2 = np(1-p) = npq. \tag{4}$$

If we use this value of σ to specify a Gaussian population,

$$y = \frac{h}{\sqrt{\pi}}e^{-h^2x^2},$$

then
$$h = \frac{1}{\sigma\sqrt{2}} = \frac{1}{\sqrt{\{2np(1-p)\}}} \quad \text{(p. 72).}$$

Bernoulli's Limit Theorem

We have seen that the mean value of the deviation $|r-np|$ is equal to $\surd(npq)$; since this expression tends to infinity with n, it follows that *when the number of trials is increased indefinitely, the probability of obtaining a deviation which is less than any assigned number tends to zero.*

At the same time we observe that the mean value of $\left|\frac{r}{n}-p\right|$ is equal to $\sqrt{\frac{pq}{n}}$, and that this expression tends to zero as n tends to infinity. We thus obtain the following fundamental result:

Theorem. *When the number n of trials is increased indefinitely, the probability that $\left|\frac{r}{n}-p\right|$ will remain less than any assigned number approaches unity.*

This theorem is due to Bernoulli, but it should be noted that the information it provides falls far short of what we should have liked to obtain. All we can infer is that the *probability* of obtaining at most a given deviation $\left|\frac{r}{n}-p\right|$ is less than any given small number, provided that n is sufficiently large—an assertion which differs essentially from the 'first empirical assumption' quoted on p. 29, from which the conception of probability has been removed.

Poisson Distributions

Bernoulli's formula states that the probability of exactly r successes in n trials is

$$P = {}^nC_r p^r(1-p)^{n-r},$$

where p is the probability of an individual event. In this formula write $p = \epsilon/n$, so that $\epsilon = np$ is, as we have seen, approximately the most probable number of successes.

Then

$$P = {}^nC_r\left(\frac{\epsilon}{n}\right)^r\left(1-\frac{\epsilon}{n}\right)^{n-r}$$

$$= \frac{n(n-1)...(n-r+1)}{r!}\left(\frac{\epsilon}{n}\right)^r\left(1-\frac{\epsilon}{n}\right)^n\Big/\left(1-\frac{\epsilon}{n}\right)^r$$

$$=\frac{\epsilon^r}{r!}\left(1-\frac{\epsilon}{n}\right)^n\frac{n(n-1)\dots(n-r+1)}{n^r}\Big/\left(1-\frac{\epsilon}{n}\right)^r$$
$$=\frac{\epsilon^r}{r!}\left(1-\frac{\epsilon}{n}\right)^n\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{r-1}{n}\right)\Big/\left(1-\frac{\epsilon}{n}\right)^r.$$

We shall now suppose that the events under consideration are rare, that is, p is small compared with unity. Hence, in order that the most probable number of successes may be appreciable, n must be large, since $p = \epsilon/n$. In these circumstances $\left(1-\frac{\epsilon}{n}\right)^n = e^{-\epsilon}$, approximately.

Now the product $\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{r-1}{n}\right)$ lies between unity and $1-\frac{r(r-1)}{2n}$, and thus tends to unity if $r(r-1)$ is small compared with $2n$, which will be the case if $r^2/2n$ is small, since r/n is still smaller.

Also $\left(1-\frac{\epsilon}{n}\right)^r = \left(1-\frac{\epsilon}{n}\right)^{n\frac{r}{n}} \to e^{-\epsilon r/n} = 1$, approximately, if n is large compared with r.

It follows that, provided $r^2/2n$ is small compared with unity and p is small, the value of P is approximately $e^{-\epsilon}\epsilon^r/r!$. In other words, if we select from a large population in which the probability p of success is small, then the probability of r successes in n trials is given by
$$P = e^{-\epsilon}\epsilon^r/r! = e^{-np}(np)^r/r!$$
provided that $r^2/2n$ in small.

This result is known as Poisson's law of distribution, applicable to the case of rare events.

Standard Deviation for Poisson's Law

The Poisson Law does not represent a true probability distribution, since the sum
$$e^{-\epsilon}\left(1+\frac{\epsilon}{1!}+\frac{\epsilon^2}{2!}+\dots+\frac{\epsilon^n}{n!}\right) = e^{-\epsilon}\sum_{r=0}^{n}\frac{\epsilon^r}{r!}$$
is not equal to unity. If, however, n is large, $\sum_{r=0}^{n}\frac{\epsilon^r}{r!}$ is approximately equal to $\sum_{r=0}^{\infty}\frac{\epsilon^r}{r!}$, by which it may be replaced.

To this degree of approximation the average value a of r is

$$a = e^{-\epsilon}\sum_{r=0}^{\infty}\frac{\epsilon^r}{r!}\times r = e^{-\epsilon}\sum_{r=0}^{\infty}\frac{\epsilon^r}{(r-1)!} = \epsilon.$$

Thus the approximate value of the standard deviation σ is given by

$$\sigma^2 = e^{-\epsilon}\sum_{r=0}^{\infty}\frac{\epsilon^r}{r!}(r-\epsilon)^2,$$

or
$$\sigma^2 = e^{-\epsilon}\left\{\epsilon^2\sum\frac{\epsilon^r}{r!}-2\epsilon\sum\frac{\epsilon^r}{(r-1)!}+\sum\frac{r\epsilon^r}{(r-1)!}\right\}$$
$$= e^{-\epsilon}\{\epsilon^2e^{\epsilon}-2\epsilon^2e^{\epsilon}+(\epsilon+\epsilon^2)e^{\epsilon}\}$$
$$= \epsilon.$$

Hence $\sigma = \sqrt{\epsilon}$.

Ex. 1. Show that the error involved in writing $\sum_{r=0}^{\infty}\frac{\epsilon^r}{r!}$ for $\sum_{r=0}^{n}\frac{\epsilon^r}{r!}$ is less than

$$\frac{e^n p^{n+1}(1-p)^{-1}}{\sqrt{\{2\pi(n+1)\}}}.$$

Ex. 2. *The Telephone Problem.* The telephone service in operation presents an enormous number of practical problems in probability. These are, however, necessarily so technical that a simple case only is given in illustration. Suppose that there are n available lines and that, on the average, ϵ of these are in operation at any given moment. Using Poisson's law we find that the probability that at any time exactly r lines are in request is

$$e^{-\epsilon}\epsilon^r/r!.$$

Now, if the average time of duration of a call is T, the probability that a call on any particular line will begin in a time dt of this interval is dt/T. Hence, the probability that a call will begin in an interval dt on any of the ϵ lines which are, on the average, in operation is $\epsilon\, dt/T$.

It follows that the probability that in the interval dt exactly r lines are in use and an additional line is required is

$$\frac{e^{-\epsilon}\epsilon^r}{r!}\cdot\frac{\epsilon\, dt}{T}.$$

Clearly, if $r \geqslant n$, the additional line will not be available and the call will be lost. Thus the probability of a call being lost on this occasion is

$$\frac{\epsilon\, dt}{T}e^{-\epsilon}\sum_{n}^{\infty}\frac{\epsilon^r}{r!}.$$

Since the probability of a call arriving in the interval dt is $\epsilon\, dt/T$, we conclude that the probable proportion of lost calls is

$$e^{-\epsilon}\sum_{n}^{\infty}\frac{\epsilon^r}{r!}.$$

Ex. 3. From a given population of N numbers $x_1, x_2,..., x_N$, a sample of magnitude n is selected. Denoting its mean by M_r, show that the mean of all the ${}^NC_n$ such means m_r is equal to the mean M of the original population.

Ex. 4. Prove that the standard deviation σ_n of the m_r's is given by

$$\sigma_n^2 = \sum (M-m_r)^2/{}^NC_n$$
$$= \frac{(N-n)}{nN^2}\sum x_i^2 - \frac{2(N-n)}{nN^2(N-1)}\sum x_i x_j.$$

Ex. 5. Deduce that the standard deviation σ of the x's is given by $\sigma^2 = \dfrac{n(N-1)}{N-n}\sigma_n^2$, so that, for large values of N, σ_n is approximately equal to $\sigma/\surd n$. (Cf. the result on p. 130, on an entirely different assumption.)

Ex. 6. Given n readings $x_1, x_2,..., x_n$ with mean m, we call the quantities $v_i = m-x_i$ the respective *residuals*. Supposing that M is the true value of m for all possible readings, we call the quantities $\epsilon_i = M-x_i$ the corresponding errors. Establish the formula $v_1 = \dfrac{n-1}{n}\epsilon_1 - \dfrac{\epsilon_2}{n} - ... - \dfrac{\epsilon_n}{n}$. Assuming now that the ϵ's are normally distributed, with precision constant h, deduce from the result of p. 129 that the precision constant h' for the v's is given by $\dfrac{1}{h'^2} = \dfrac{n-1}{nh^2}$, and hence that the standard deviation for the ϵ's is $\surd\{\sum v_i^2/(n-1)\}$.

CHAPTER IX

THE USE OF PROBABILITY IN SCIENTIFIC INDUCTION

1. The general problem

All scientific conclusions are arrived at by a combination of inductive and deductive processes. The experimenter provides the data, the mathematician accepts them and offers a hypothesis which links them together, and then by mathematically deductive reasoning draws certain conclusions from them. From the point of view of mathematical technique a deduction has been made; from that of scientific method, in stating a hypothesis which outruns the experimental data alone, an induction is involved. The mathematician has deduced certain consequences, and, offering them to the experimenter as possible truths, demands their physical verification or disproof. The experimenter deduces by his particular method that they are true in his particular circumstances; and together they pass to the inductive stage that the hypothesis outstripping even these new facts is still true in the sense that it is a valid guide to the next step.

Thus we discover three elements in any scientific problem:

(1) A set of data, given as the result of experiment: we refer to these as the 'sample'.

(2) A wider field ('the population') of possible data from which (1) has been selected.†

(3) A hypothesis or hypothetical law tentatively presumed to govern the structure of (2).

Stated in this way, the problem appears in a form detached from the experimental methods which are necessary to collect the data and from the use to which (3) is to be put. For example, on account of the imperfections of their apparatus the experimenters may incorporate in a reading at time t, say, readings over a time $t \pm t'$. Or the data may be such as to require classification as of length l when in fact the actual

† In this connexion see the limitations of this principle in many cases of physical science (Ch. VIII).

lengths vary about l. It follows that there is apparently a fourth factor in the situation which requires to be considered if (1), (2), and (3) are to appear as steps in the scientific process, namely,

(4) The process of selecting or 'sampling' the data.

The way in which these four elements are associated can be shown in mathematical form. Let us imagine an original population to consist entirely of elements having a common characteristic measured by the variable t; and suppose that this characteristic occurs at values t_1, t_2,... with frequencies $V(t_1)$, $V(t_2)$,..., where for the moment t_1, t_2,... are integers. The total size of the population is thus

$$V(t_1)+V(t_2)+...+V(t_n).$$

In a problem of scientific induction we do not know the form of $V(t)$; we can only speculate on it by means of (1) and (3). Suppose, however, that a set of data $U(t)$ has been collected, covering the whole range of t: thus, $U(t_r)$ is the frequency with which the data are collected at what the experimenter believes to be the value t_r. We have used the word 'believes' designedly because what the experimenter does in practice is to sweep into his reading at, say, t_1 a number of readings at $t_1 \pm 1$, $t_1 \pm 2$, etc.; this inclusion of false data is not within his control, for he acts on the assumption that he is obtaining correct data at the given value of t.

We shall suppose that the false data are swept into the true readings according to some particular law; thus, let the unknown law which describes the proportions of readings at neighbouring positions included in the reading at t be $p(s)$, where s is the interval between the readings at t and $t+s$. Since $V(t+s)$ is the number of readings which occur at $t+s$, the number of these which are accepted as being at t is $V(t+s)p(s)$.

It follows that the frequency $U(t)$ of the samples found at t is the sum of all terms of the type $V(t+s)p(s)$, where s takes all possible values about the position t. It is clear that a good experimenter will have so designed his experiment that very few, and small, values of s occur; mathematically this implies simply that $p(s)$ is always zero beyond a particular range of s.

With this understanding we can write

$$U(t) = \sum_{s=-\infty}^{\infty} V(t+s)p(s). \tag{1}$$

Ex. 1. Suppose that the sample is obtained by including with half the values which actually occur at t, a quarter of those which occur on both sides. Then the function $p(s)$ is defined by the properties

$$p(0) = \tfrac{1}{2}, \qquad p(1) = p(-1) = \tfrac{1}{4}, \quad \text{and} \quad p(s) = 0$$

for all other values of s. It follows from (1) that

$$U(t) = V(t-1)p(-1)+V(t)p(0)+V(t+1)p(1),$$

i.e. $$2U(t) = V(t)+\tfrac{1}{2}[V(t-1)+V(t+1)].$$

Thus, if, for example, $V(t) = t(10-t)$, then

$$2U(t) = 20t-2t^2-1.$$

Ex. 2. If $V(t)$ is given by the table

t	0	1	2	3	4	5	6	7
$V(t)$	1	3	7	11	14	11	7	3

calculate the nature of the sample $U(t)$ for values of t from $t = 1$ to $t = 6$, using the method of selection in Ex. 1.

The above examples illustrate the simple problem of deducing the sample when the structure of the original population and the mode of selection are specified.

We are now in a position to restate our previous remarks in symbolical form. We are confronted with the following problem: If a sample distribution $U(t)$ has been found and a method of selection $p(s)$ postulated, what can be deduced about the original population $V(t)$?

If the operator E is defined by the relation

$$Ef(t) = f(t+1),$$

so that $$E^s f(t) = f(t+s),$$

then (1) becomes

$$U(t) = \sum_{-\infty}^{\infty} V(t+s)p(s) = \sum_{-\infty}^{\infty} E^s V(t)p(s)$$
$$= \Big\{\sum_{-\infty}^{\infty} E^s p(s)\Big\} V(t).$$

If the infinite series within the brackets has the formal sum $\phi(E)$, then we obtain†

$$U(t) = \phi(E)V(t),$$

† Cf. Chapter II, p. 29.

whence, by the method of operators, the particular solution of this equation is

$$V(t) = \phi^{-1}(E)U(t). \tag{2}$$

Now

$$\begin{aligned} Ef(t) &= f(t+1)-f(t)+f(t) = \Delta f(t)+f(t), \\ &= (\Delta+1)f(t), \end{aligned}$$

in the notation of differences.

Suppose that $\phi^{-1}(E) \equiv \phi^{-1}(1+\Delta)$ can be expanded in ascending powers of Δ, in the form

$$A_0+A_1\Delta+A_2\Delta^2+\dots.$$

Then the function $V(t)$ which represents the original population is expressed in terms of the sample $U(t)$ and its differences. We note that if $U(t)$ can be represented as a polynomial in t, then all its differences beyond a certain power are zero and $V(t)$ is expressed in finite terms.

Ex. 3. Consider the equation given in Ex. 1 above. We have

$$2U(t) = \{\tfrac{1}{2}(E^{-1}+E)+1\}V(t) = \frac{E^2+2E+1}{2E}V(t).$$

Thus

$$\begin{aligned} V(t) &= \frac{4E}{(E+1)^2}U(t) = \frac{4(1+\Delta)}{(2+\Delta)^2}U(t) \\ &= (1+\Delta)(1+\tfrac{1}{2}\Delta)^{-2}U(t) \\ &= (1+\Delta)(1-\Delta+\tfrac{3}{4}\Delta^2\dots)U(t) \\ &= (1-\tfrac{1}{4}\Delta^2\dots)U(t). \end{aligned}$$

If, for example, $U(t) = 9-t^2$ in the range $(-3, 3)$, then

$$V(t) = 9-t^2+\tfrac{1}{2}.$$

Ex. 4. Suppose that $p(s) = e^{-s}$ $(s > 0)$ and that $p(s) = 0$ $(s < 0)$.

Then

$$U(t) = \sum_0^\infty \left(\frac{E}{e}\right)^s V(t) = \frac{e}{e-E}V(t).$$

Hence

$$V(t) = \frac{e-E}{e}U(t) = U(t)-\frac{1}{e}U(t+1).$$

We return now to consider the general solution of equation (1). This consists of the particular solution (2) and a 'complementary function', the solution of

$$\phi(E)V(t) = 0.$$

This function is to some extent arbitrary in character, as is seen by the following examples.

Ex. 1. Suppose that $U(t) = 15-t^2$, and that the sample is obtained from the population $V(t)$ by the law

$$U(t) = \tfrac{1}{2}[V(t+1)+V(t-1)],$$

where the range of values of t required for the evaluation of $U(t)$ is given by
$$-2 \leqslant t \leqslant 2.$$

We have to solve the equation
$$(E^2+1)V(t-1) = 2(15-t^2).$$
We thus obtain for the general solution
$$V(t) = 16-t^2+A\cos\tfrac{1}{2}\pi t+B\sin\tfrac{1}{2}\pi t,$$
where A and B are arbitrary constants or functions of period unity.

Now $V(t)$ must remain positive over the whole range of t required, namely $-3 \leqslant t \leqslant 3$. We shall see that this condition may be secured by taking $A = 0$; for in that case, $V(t)$ will remain positive in the required range, provided that B satisfies the condition
$$-7 \leqslant B \leqslant 7.$$
Hence there is an infinity of solutions to our equation satisfying the given conditions for $U(t)$.

Ex. 2. That a hypothetical population cannot always be found may be seen from the following example. Suppose instead that $U(t) = 16-t^2$, and that the law of selection is the same as before. Since $U(t)$ must be positive, we require $-4 \leqslant t \leqslant 4$. The general solution of the equation for $V(t)$ is found to be
$$V(t) = 17-t^2+A\cos\tfrac{1}{2}\pi t+B\sin\tfrac{1}{2}\pi t.$$

With our law of selection $V(t)$ must certainly be positive in the range $-5 \leqslant t \leqslant 5$. But, substituting $t = 5$ and $t = -5$ in the solution, this necessitates $B > 8$ and $B < -8$, which is impossible. It follows that, *with the given law of selection, no population can be found to yield the given sample.*

Ex. 3. If $3U(t) = V(t-1)+V(t)+V(t+1)$, then
$$3U(t) = \frac{(E^2+E+1)}{E}V(t).$$

The complementary function is evidently
$$V(t) = A\omega_1^t+B\omega_2^t, \tag{1}$$
where A and B are arbitrary and ω_1, ω_2 are the roots of the equation
$$E^2+E+1 = 0,$$
i.e. the complex cube roots of unity.

If we write $\omega_1 = \cos\tfrac{2}{3}\pi+i\sin\tfrac{2}{3}\pi$, $\omega_2 = \cos\tfrac{2}{3}\pi-i\sin\tfrac{2}{3}\pi$, (1) may be expressed in the form
$$V(t) = A\cos(\tfrac{2}{3}\pi t+\alpha),$$
where A and α are arbitrary constants.

The particular solution is given by

$$V(t) = \frac{3E}{E^2+E+1}U(t) = \frac{3(1+\Delta)}{(1+\Delta)^2+(1+\Delta)+1}U(t),$$

so that

$$\begin{aligned} V(t) &= (1+\Delta)(1+\Delta+\tfrac{1}{3}\Delta^2)^{-1}\,U(t) \\ &= (1+\Delta)(1-\Delta-\tfrac{1}{3}\Delta^2+\Delta^2...)\,U(t) \\ &= (1-\tfrac{1}{3}\Delta^2)\,U(t) \qquad (2) \end{aligned}$$

if higher differences of $U(t)$ may be neglected.

To this order of approximation, the general solution of the equation is

$$V(t) = A\cos(\tfrac{2}{3}\pi t+\alpha)+U(t)-\tfrac{1}{3}\Delta^2U(t). \qquad (3)$$

It is clear that the determination of the hypothetical population (3) above is equivalent to the process of graduating or 'smoothing' the errors introduced by the selective process, as is explained in a later section. Our sample $U(t)$ has been formed by taking the mean of three adjacent ordinates of the histogram $V(t)$, and our solution (3) represents analytically a reversal of this process. If we confine our attention to the particular solution, for which $A = 0$, we note that when $U(t)$ is a linear function of t, $\Delta^2U(t)$ is zero, so that $V(t) = U(t)$. When $U(t)$ is a quadratic function of t, $V(t)$ and $U(t)$ differ only by a constant.

Ex. 4. Find the original population $V(t)$, given that each reading shown for $U(t)$ is the true reading at t plus $\frac{1}{10}$th of the true reading at $t+1$.

t	0	1	2	3	4	5	6	7	8	9
$U(t)$	13	22·8	31·2	35·5	38·6	39·5	38·2	34·8	30	21·1

We have to solve the equation

$$U(t) = V(t)+\tfrac{1}{10}V(t+1).$$

For the readings shown this gives the solution

$$V(t) = 36-(t-5)^2.$$

Bernoullian Law of Selection

Consider first the case in which $p(s) = {}^2C_{s+1}p^{s+1}(1-p)^{1-s}$, where p is a given positive fraction. The equation (1), p. 148, then becomes

$$U(t) = (1-p)^2V(t-1)+2p(1-p)V(t)+p^2V(t+1).$$

We have thus applied a Bernoulli process of selection to the set of three consecutive ordinates of the histogram $V(t)$ in order to obtain the sample $U(t)$; and if $\frac{1}{3} < p < \frac{2}{3}$, the ordinate at t is swept into the readings with a greater probability than either of the adjacent ordinates. If we put $q = 1-p$, the equation may be written symbolically as

$$U(t) = (q^2/E+2qp+p^2E)V(t),$$

or
$$U(t) = \frac{(pE+q)^2}{E}V(t).$$

Thus the particular solution is given by

$$V(t) = \frac{E}{(pE+q)^2}U(t) = \{1+\Delta\}\{1+p\Delta\}^{-2}U(t)$$
$$= (1+\Delta)[1-2p\Delta+3p^2\Delta^2-...+(-1)^{n-1}np^{n-1}\Delta^{n-1}]U(t)$$
$$= 1+\sum_{n=1}^{\infty}(-1)^n p^{n-1}[(n+1)p-n]\Delta^n U(t).$$

Suppose generally that $p(s)$ is defined by the formula

$$p(s) = {}^nC_{s+m}\,p^{s+m}q^{n-m-s},$$

where n and m are given numbers. There will now be $n+1$ terms on the right-hand side of (1), which becomes

$$U(t) = \sum_{s=-m}^{n-m} {}^nC_{s+m}\,p^{s+m}q^{n-m-s}V(t+s),$$

or, symbolically,
$$U(t) = \frac{(pE+q)^n}{E^m}V(t).$$

Hence the particular solution is

$$V(t) = E^m(1+p\Delta)^{-n}U(t) = (1+p\Delta)^{-n}U(t+m).$$

The general solution is thus

$$V(t) = (1+p\Delta)^{-n}U(t+m)+\left(\frac{p-1}{p}\right)^t\{A_1+A_2t+...+A_{n-1}t^{n-1}\},$$

where A_1, A_2,... are arbitrary constants or functions of period unity.

Bayes's Theorem

Bayes's theorem, which by its misapplication has attained a certain notoriety in the history of probability, follows at once from the foregoing discussion. In Fig. 25 the population $V(t)$, from which the sample $U(t)$ is drawn, may be regarded as contributing its quota to the sample at t in the proportions indicated. As we have seen, the total sample is

$$U(t) = \sum_{s=-\infty}^{\infty} V(t+s)p(s).$$

The contribution to this total at any position distant s from t is $V(t+s)p(s)$. Thus, given a full knowledge of the population $V(t)$ and the process $p(s)$ of selection, we can say that *a member of the sample at t has a probability* $\dfrac{V(t+s_1)p(s_1)}{\sum\limits_{-\infty}^{\infty} V(t+s)p(s)}$ *that it has come from the position* $t+s_1$ *in the* $V(t)$ *diagram.*

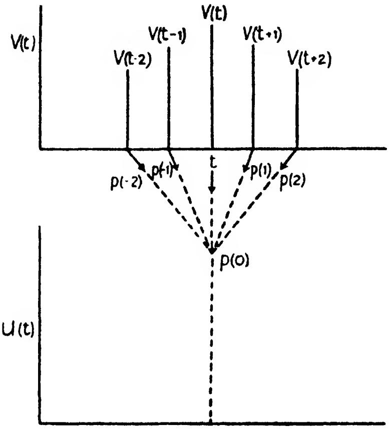


FIG. 25

This, in effect, is Bayes's theorem. The frequency function $U(t)$ enables us to specify the probability that a member of the sample will lie at t; this is the initial probability conditioned only by the statement that the individual is a member of $U(t)$. At this stage the theorem enters to tell us the probable source of this value of t when further information is available—the information being that the distribution $U(t)$ has been derived from the source $V(t)$ by a certain process $p(s)$.

Ex. 1. Three boxes contain balls as shown:

Box 1	*Box* 2	*Box* 3
1 black 1 white	1 black 3 yellow	1 black 4 green

It is known that a fourth box has dropped into it a ball from Box 1, two balls from Box 2, and one from Box 3. What is the probability that a ball in this box, known to be black, came from Box 2?

Without the information that it is black, the probability that it came from Box 2 is clearly $\frac{2}{1+2+1} = \frac{1}{2}$. We ask, what difference is made in this probability by the additional information that the ball is black. Actually, the additional fact converts the problem into a new one; and the comparison of the answers to the two problems, a step usually associated with Bayes's theorem, has nothing to do with the question.

In order to calculate the required probability we have to construct the functions $V(t)$ and $p(s)$, the variable t being the suffix attached to each of the three boxes. Thus, $V(1)$, $V(2)$, and $V(3)$ are the numbers of balls in Box 4 which come from Boxes 1, 2, and 3, respectively. The probabilities of a black ball in the three cases are

$$p(1) = \tfrac{1}{2},\quad p(2) = \tfrac{1}{4},\quad \text{and}\quad p(3) = \tfrac{1}{5}, \text{ respectively.}$$

Hence $$\sum V(t+s)p(s) = 1.\tfrac{1}{2}+2.\tfrac{1}{4}+1.\tfrac{1}{5} = \tfrac{6}{5}.$$

The contribution $V(2)p(2)$ is $2.\frac{1}{4}$. Thus the required probability is $\frac{1}{2}/\frac{6}{5} = \frac{5}{12}$.

The distinction between the two problems is now clear: the probability of a ball drawn from Box 4 having come from Box 2 is $\frac{1}{2}$, while if a ball is drawn from Box 4 and found to be black, the probability that it came from Box 2 is $\frac{5}{12}$.

Ex. 2. Given n_1 urns A_1 each containing ν_1 white balls, n_2 urns A_2 each containing ν_2 white balls,... and n_r urns A_r each containing ν_r white balls: one of the urns is chosen and a ball extracted, which turns out to be white. What is the probability that it came from one of the n_1 urns A_1?

We may suppose the balls placed together in one urn, provided it is always possible to specify the urns from which they came: we do not thus alter the probability of extracting a white ball. The total number of white balls is $n_1\nu_1+n_2\nu_2+...+n_r\nu_r$, of which $n_1\nu_1$ come from the urns A_1. If now a white ball is extracted, the probability that it is from one of the set A_1 is

$$n_1\nu_1/(n_1\nu_1+n_2\nu_2+...+n_r\nu_r).$$

In applications of Bayes's theorem, it must be understood that the structure of the original population is precisely delimited: what we ask is whether, when a particular event among a series occurs, its source can be traced to this or that element of the structure. When the problem is stated in this way, Bayes's theorem gives a definite answer. If, however, we attempt to recast the problem by raising a query about the *structure* of the population, then we are faced with the solution of an equation (p. 148) to which a unique answer cannot necessarily be given, since an element of arbitrariness is present.

Ex. 3. An urn contains a black and white balls, in unknown proportions: a ball is extracted n times and each time replaced in the urn. If ν of the balls extracted are white, what is the probability that α of the balls in the urn are white?

The required probability is that of a subclass of the subclass of urns in the population of urns containing a black and white balls, which contain precisely α white balls. Thus we must imagine the urn in question to come from a population of urns, each of which contains a black and white balls, the population covering all possible compositions. In this population, the first subclass consists of urns containing no white balls, the second consists of urns containing one white ball, and so on. Then the probability that, if an urn containing i white balls is selected, ν white balls will be obtained in n extractions, is by Bernoulli's theorem,

$$^{n}C_{\nu}\left(\frac{i}{a}\right)^{\nu}\left(1-\frac{i}{a}\right)^{n-\nu}.$$

Now suppose that the probabilities of choosing the first, second, third,... subclasses of urn are p_0, p_1, p_2,..., respectively. Then, by Bayes's theorem, the probability that the urn chosen is one containing α white balls is

$$\frac{\left\{^{n}C_{\nu}\left(\frac{\alpha}{a}\right)^{\nu}\left(1-\frac{\alpha}{a}\right)^{n-\nu}p_{\alpha}\right\}}{\left\{^{n}C_{\nu}\left(\frac{1}{a}\right)^{\nu}\left(1-\frac{1}{a}\right)^{n-\nu}p_1+{}^{n}C_{\nu}\left(\frac{2}{a}\right)^{\nu}\left(1-\frac{2}{a}\right)^{n-\nu}p_2+\ldots\right\}}$$

$$=\frac{\alpha^{\nu}(a-\alpha)^{n-\nu}p_{\alpha}}{\{1^{\nu}(a-1)^{n-\nu}p_1+2^{\nu}(a-2)^{n-\nu}p_2+\ldots+(a-1)^{\nu}1^{n-\nu}p_{a-1}\}}.$$

It will be noted, therefore, that the solution of the problem depends on a knowledge of the probabilities p_1, p_2,... about which we have no information whatever. If we make the *assumption* that all types of urn have equal probability, then $p_1 = p_2 = ... = p_{a-1}$, and the probability sought is

$$\alpha^{\nu}(a-\alpha)^{n-\nu}/\{1^{\nu}(a-1)^{n-\nu}+2^{\nu}(a-2)^{n-\nu}+...+(a-1)^{\nu}1^{n-\nu}\}.$$

It is readily shown that this is a maximum for the value of α such that $\frac{\alpha}{a} = \frac{\nu}{n}$: that is, the most probable composition of the urn is that which gives for the required probability the value $\frac{\nu}{n}$.

Extension to Functions of a Continuous Variable

Let $V(t)$ be a function of a continuous variable t which is defined in the range $(-\infty, \infty)$ and which gives the probability $V(t)$ of the occurrence of the variable in the interval dt about the position t. Suppose that a new population is constructed from the distribution according to the following law: at a distance x from the position t, the ordinate $V(t+x)$ is to be swept in with a probability $p(x)$ and allocated to the position t, the value of x extending over the range $a \leqslant x \leqslant b$. If $U(t)$ is the probability, in the new population, of a value t occurring in an interval dt about t, then the contribution to $U(t)$ at the position t is given by

$$V(t+x)p(x).$$

Thus the probability $U(t)$ is given by

$$U(t) = \int\limits_a^b V(t+x)p(x)\, dx.$$

It should be noticed that if the original probability function $V(t)$ has a *finite* range, then the function $V(t+x)$, for values of x which take it beyond this range, is, of course, zero and makes no contribution to the integral. Thus $U(t)$ has exactly the same range as the original function $V(t)$.

Ex. 1. An interesting application of the previous results has been made by Eddington.† Suppose that the probability function $u(t)$ for the sample is given, and that the law of selection is Gaussian, i.e.

$$p(x) = \frac{h}{\sqrt{\pi}}\exp(-h^2x^2).$$

† Eddington, *Monthly Notices, R.A.S.* **73**, 359.

Then the probability function $v(t)$ of the original population is given by the equation

$$u(t) = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} v(t+x)\exp(-h^2x^2)\,dx.$$

Now write

$$\begin{aligned} v(t+x) &= v(t)+x\frac{dv}{dt}+\frac{x^2}{2!}\frac{d^2v}{dt^2}+\dots \\ &= \left(1+x\frac{d}{dt}+\frac{x^2}{2!}\frac{d^2}{dt^2}+\dots\right)v(t) \\ &= \exp\left(x\frac{d}{dt}\right)v(t),\text{ symbolically.} \end{aligned}$$

Thus $$u(t) = \left[\frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(x\frac{d}{dt}-h^2x^2\right)dx\right] v(t).$$

The integral is an operator as regards t, but a definite integral as regards x.

Now $$\int_{-\infty}^{\infty} \exp(-ax-bx^2)\,dx = \sqrt{\frac{\pi}{b}}\exp(a^2/4b);$$

then writing $a = -\frac{d}{dt}$, $b = h^2$, we have

$$u(t) = \frac{h}{\sqrt{\pi}}\left\{\sqrt{\frac{\pi}{h^2}}\exp\left(\frac{d^2}{dt^2}\Big/4h^2\right)\right\}v(t).$$

Hence

$$\begin{aligned} v(t) &= \exp\left(-\frac{d^2}{dt^2}\Big/4h^2\right)u(t) \\ &= \left\{1-\frac{1}{4h^2}\frac{d^2}{dt^2}+\frac{1}{2!}\left(\frac{1}{4h^2}\right)^2\frac{d^4}{dt^4}-\dots\right\}u(t) \\ &= u(t)-\frac{1}{4h^2}u''(t)+\frac{1}{2(4h^2)^2}u^{\text{iv}}(t)-\dots. \end{aligned}$$

When h is large, it is sufficient to consider the first few terms of this expression. Since $u(t)$ is an empirical probability function, it is better to express $v(t)$ in terms of $u(t)$ and its successive differences rather than its differential coefficients.

Now $$u''(t) = \Delta^2u(t)-\Delta^3u(t)+\frac{11}{12}\Delta^4u(t)-\dots,$$

and $$u^{\text{iv}}(t) = \Delta^4u(t)+\dots.$$

Hence

$$v(t) = u(t)-\frac{1}{4h^2}\left\{\Delta^2-\Delta^3+\frac{11}{12}\Delta^4\right\}u(t)+\frac{1}{2(4h^2)^2}\Delta^4u(t)...+$$

$$= u(t)-\frac{1}{4h^2}\Delta^2u(t)+\frac{1}{4h^2}\Delta^3u(t)-\frac{1}{4h^2}\left(\frac{11}{12}-\frac{1}{8h^2}\right)\Delta^4u(t)+....$$

A comparison between the result just obtained and Bayes's theorem is inevitable. There, the passage back from the function $U(t)$ to $V(t)$ was not necessarily possible for any given function $U(t)$ nor, when possible, was it necessarily unique; we saw that this circumstance was associated with the fact that the range of $U(t)$ was not in general that of $V(t)$, and that this depended entirely on the law of selection $p(x)$. In our example, however, the variable t is continuous and the two ranges are identical; the passage back, when it can be achieved, is unique—no arbitrariness is involved. In that case, Bayes's theorem, as stated above, gives the probability $U(t)$ that a certain variable t, derived from a function $V(t)$ calculated by the process defined by $p(x)$, came from the range (a, b). This is the inverse form of Bayes's theorem as usually applied to determine the 'probability of causes', and the application is legitimate if we bear in mind that the method of selection $p(x)$ is assumed to be given; it cannot be chosen arbitrarily.

From a knowledge of $U(t)$ both $V(t)$ *and* $p(x)$ cannot be determined separately; and it is by ignoring this vital fact and by tacitly assuming that $p(x)$ is some such function as unity or $\frac{h}{\sqrt{\pi}}\exp(-h^2x^2)$, that writers have been led to conclude that Bayes's theorem may be used to trace back, with a certain degree of probability, the antecedent events which have given rise to the function $U(t)$. This procedure, as we have seen, is wholly fallacious.

In the foregoing example it is assumed that the law of selection is the normal error law. If this law is not obeyed in the case to which it is applied, the result will be invalid in practice.

Ex. 2. Let us now suppose that the functions $v(x)$ and $p(x)$ are both Gaussian, so that

$$v(x) = \frac{h}{\sqrt{\pi}}\exp(-h^2x^2), \text{ and } p(x) = \frac{h'}{\sqrt{\pi}}\exp(-h'^2x^2),$$

where h and h' are known constants. Then $u(t)$ is given by

$$u(t) = \frac{hh'}{\pi}\int\limits_{-\infty}^{\infty} \exp\{-h^2(t+x)^2-h'^2x^2\}\,dx$$

$$= \frac{hh'}{\pi}\exp\left(-\frac{h^2h'^2t^2}{h^2+h'^2}\right)\int\limits_{-\infty}^{\infty} \exp\left\{-(h^2+h'^2)\left(x+\frac{h^2t}{h^2+h'^2}\right)^2\right\}dx,$$

by completing the square in the exponent.

By changing the variable to $z = x+\frac{h^2t}{h^2+h'^2}$, we reduce the integral to the form

$$\int\limits_{-\infty}^{\infty} \exp\{-(h^2+h'^2)z^2\}\,dz = \frac{\sqrt{\pi}}{\sqrt{(h^2+h'^2)}}$$

on evaluation.

It follows that

$$u(t) = \frac{hh'}{\sqrt{\pi}\sqrt{(h^2+h'^2)}}\exp\left(-\frac{h^2h'^2}{h^2+h'^2}t^2\right).$$

Hence the function $u(t)$ also follows a Gaussian law

$$u(t) = \frac{h''}{\sqrt{\pi}}\exp(-h''^2t^2),$$

where $$h''^2 = \frac{h^2h'^2}{h^2+h'^2}.$$

If the standard deviations of $v(x)$ and $p(x)$ are σ and σ', respectively, so that $\sigma = \frac{1}{h\sqrt{2}}$, $\sigma' = \frac{1}{h'\sqrt{2}}$, the standard deviation σ'' of $u(t)$ is therefore given by $\sigma''^2 = \sigma^2+\sigma'^2$.

Evidently the theorem we have obtained can be inverted; for if $u(t)$ and $p(x)$ are both Gaussian functions, similar reasoning shows that $v(x)$ must also be Gaussian. Thus, in conclusion, we have the result:

If the distribution of the original population and the probability of sampling follow the Gaussian law, then the sample also follows the Gaussian law; and if the sample and the probability of sampling follow the Gaussian law, so does the original population.

Ex. 3. This result may be extended to a series of samples, each of which is drawn from the preceding one. Thus, suppose that $v(t)$ is a Gaussian population and that $\exp(-h^2x^2)$ is the probability of the sampling. Then a sample $u_1(t)$ of the population is given by

$$u_1(t) = \int\limits_{-\infty}^{\infty} v(t+x)\exp(-h^2x^2)\,dx.$$

Then, by the above theorem, $u_1(t)$ is of the form $A\exp(-h_1^2t^2)$, where

$$\frac{1}{h_1^2} = \frac{1}{h^2}+\frac{1}{h'^2}.$$

If now a sample $u_2(t)$ is drawn from the sample $u_1(t)$, its magnitude is given by

$$u_2(t) = \int\limits_{-\infty}^{\infty} u_1(t+x)\exp(-h'^2x^2)\,dx,$$

and the corresponding constant h_2 satisfies the relation

$$\frac{1}{h_2^2} = \frac{1}{h_1^2}+\frac{1}{h'^2}.$$

Similarly for a sample $u_3(t)$ drawn from $u_2(t)$, and so on. Hence the constant h_n specifying the nth sample in this succession is given by

$$\frac{1}{h_n^2} = \frac{1}{h_{n-1}^2}+\frac{1}{h'^2}.$$

Adding the n equations so obtained, we have

$$\frac{1}{h_n^2} = \frac{1}{h^2}+\frac{n}{h'^2}.$$

In terms of the standard deviations this becomes†

$$\sigma_n^2 = \sigma^2+n\sigma'^2.$$

Two-dimensional Distributions

Suppose, for instance, that a sample $U(x,y)$ is obtained by taking the mean of the values of the population $V(x,y)$ at the four points $(x\pm 1, y\pm 1)$. We then have the equation

$$4U(x,y) = V(x+1,y+1)+V(x-1,y+1)+ \\ +V(x-1,y-1)+V(x+1,y-1).$$

† Cf. p. 130.

If we write

$$EV(x,y) = V(x+1,y)$$

and

$$FV(x,y) = V(x,y+1),$$

we obtain

$$\begin{aligned}4U(x,y) &= (EF+E^{-1}F+E^{-1}F^{-1}+EF^{-1})V(x,y)\\ &= (E+E^{-1})(F+F^{-1})V(x,y)\\ &= \frac{(E^2+1)(F^2+1)}{EF}V(x,y).\end{aligned}$$

The particular solution of this equation is

$$\begin{aligned}V(x,y) &= \frac{4EF}{(E^2+1)(F^2+1)}U(x,y)\\ &= \frac{(1+\Delta)(1+\Delta')}{\{1+\frac{1}{2}(2\Delta+\Delta^2)\}\{1+\frac{1}{2}(2\Delta'+\Delta'^2)\}}U(x,y),\end{aligned}$$

where the operators Δ and Δ' refer to x and y respectively.

Thus

$$\begin{aligned}V(x,y) &= (1+\Delta)(1+\Delta')(1-\Delta+\tfrac{1}{2}\Delta^2-...)\times\\ &\qquad\qquad\times(1-\Delta'+\tfrac{1}{2}\Delta'^2-...)U(x,y)\\ &= (1-\tfrac{1}{2}\Delta^2+...)(1-\tfrac{1}{2}\Delta'^2+...)U(x,y)\\ &= U(x,y)-\tfrac{1}{2}(\Delta^2+\Delta'^2)U(x,y)+\tfrac{1}{4}\Delta^2\Delta'^2U(x,y)....\end{aligned}$$

To this must be added the complementary function

$$A\sin\tfrac{1}{2}\pi x+B\cos\tfrac{1}{2}\pi x+C\sin\tfrac{1}{2}\pi y+D\cos\tfrac{1}{2}\pi y.$$

Ex. A certain substance is being deposited on the inside of a tube 6 cm. in length, and measurements of the extent of the deposit are taken at intervals of one second at distances of 1 cm. along the tube. The following are the results obtained (in grammes):

		Values of t				
		0	1	2	3	4
Values of x	1	0·35	0·45	0·55	0·65	0·75
	2	0·65	0·85	1·05	1·25	1·45
	3	1·15	1·45	1·75	2·05	2·35
	4	1·85	2·25	2·65	3·05	3·45
	5	2·75	3·25	3·75	4·25	4·75
	6	3·85	4·45	5·05	5·65	6·25

Assuming that the readings at (x,t) were really the average of those at x and $x+1$, taken at times t and $t+1$ respectively, correct the above data so as to give the true values at (x,t).

If the true value is $V(x,t)$ and the sample is $U(x,t)$, then

$$U(x,t) = \tfrac{1}{2}\{V(x,t)+V(x+1,t+1)\}$$
$$= \tfrac{1}{2}(EF+1)V(x,t).$$

Hence

$$V(x,t) = \frac{2}{EF+1}U(x,t)$$
$$= \frac{2}{(1+\Delta)(1+\Delta')+1}U(x,t)$$
$$= \{1+\tfrac{1}{2}(\Delta+\Delta'+\Delta\Delta')\}^{-1}U(x,t)$$
$$= \{1-\tfrac{1}{2}(\Delta+\Delta'+\Delta\Delta')+\tfrac{1}{4}(\Delta+\Delta'+\Delta\Delta')^{2}-...\}U(x,t),$$

and the problem is reduced to that of constructing a twofold difference table from the one given.

Two-dimensional Continuous Distributions

Suppose that the given sample $u(x,y)$ is a continuous function of two independent variables x, y, and that it is derived from a population $v(x,y)$ which is also a continuous function of the same variables. Let the selective process $p(\xi,\eta)$ by which $u(x,y)$ is obtained from $v(x,y)$ be such that the probability of choosing a sample in a region of area $d\xi d\eta$ surrounding the point $(x+\xi, y+\eta)$ is $p(\xi,\eta)\,d\xi d\eta$. Then the law connecting sample and population is evidently

$$u(x,y) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} v(x+\xi,y+\eta)p(\xi,\eta)\,d\xi d\eta.$$

Let us apply this result to the case in which the law of selection is Gaussian, so that, for instance,

$$p(\xi,\eta) = \frac{hk}{\pi}\exp(-h^{2}\xi^{2}-k^{2}\eta^{2}).$$

We then have

$$u(x,y) = \frac{hk}{\pi}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} v(x+\xi,y+\eta)\exp(-h^{2}\xi^{2}-k^{2}\eta^{2})\,d\xi d\eta.$$

If we denote the operators $\partial/\partial\xi$, $\partial/\partial\eta$ by D and D' respectively, we may write

$$v(x+\xi,y+\eta) = v(x,y)+(\xi D+\eta D')v(x,y)+\frac{(\xi D+\eta D')^{2}}{2!}v(x,y)...$$
$$= \exp(\xi D+\eta D')v(x,y).$$

Hence, symbolically,

$$u(x,y) = \frac{hk}{\pi}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \exp(\xi D+\eta D')v(x,y)\exp\{-(h^2\xi^2+k^2\eta^2)\}\,d\xi d\eta$$

$$= \frac{hk}{\pi}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \exp(\xi D-h^2\xi^2)\exp(\eta D'-k^2\eta^2)v(x,y)\,d\xi d\eta$$

$$= \frac{hk}{\pi}\int\limits_{-\infty}^{\infty} \exp(\xi D-h^2\xi^2)\,d\xi \int\limits_{-\infty}^{\infty} \exp(\eta D'-k^2\eta^2)\,d\eta\; v(x,y).$$

Now
$$\frac{h}{\sqrt{\pi}}\int\limits_{-\infty}^{\infty} \exp(\xi D-h^2\xi^2)\,d\xi = \exp(D^2/4h^2)$$

and
$$\frac{k}{\sqrt{\pi}}\int\limits_{-\infty}^{\infty} \exp(\eta D'-k^2\eta^2)\,d\eta = \exp(D'^2/4k^2),$$

so that
$$u(x,y) = \exp(D^2/4h^2+D'^2/4k^2)v(x,y).$$

Thus
$$v(x,y) = \exp\left\{-\frac{1}{4}\left(\frac{D^2}{h^2}+\frac{D'^2}{k^2}\right)\right\}u(x,y),$$

or
$$v(x,y) = \left[1-\frac{1}{4h^2}D^2+\frac{1}{2!}\frac{1}{(4h^2)^2}D^4-...\right]\times$$
$$\times\left[1-\frac{1}{4k^2}D'^2+\frac{1}{2!}\frac{1}{(4k^2)^2}D'^4-...\right]u(x,y).$$

Ex. Suppose that $u(x,y)$ is the mean value over a square of side $2T$ about the point (x,y); then

$$u(x,y) = \frac{1}{4T^2}\int\limits_{-T}^{T}\int\limits_{-T}^{T} v(x+\xi,y+\eta)\,d\xi d\eta$$

$$= \frac{1}{4T^2}\int\limits_{-T}^{T}\int\limits_{-T}^{T} \exp(\xi D+\eta D')v(x,y)\,d\xi d\eta, \quad \text{as before,}$$

$$= \frac{1}{4T^2}\int\limits_{-T}^{T} \exp(\xi D)\,d\xi \int\limits_{-T}^{T} \exp(\eta D')\,d\eta\; v(x,y)$$

$$= \frac{1}{4T^2}\left[\frac{\exp(TD)-\exp(-TD)}{D}\right]\left[\frac{\exp(TD')-\exp(-TD')}{D'}\right]v(x,y)$$

$$= \frac{1}{T^2}\frac{\sinh TD \sinh TD'}{DD'}v(x,y).$$

Hence

$$v(x,y) = \frac{TD}{\sinh TD}\,\frac{TD'}{\sinh TD'}\,u(x,y)$$

$$= (1+\tfrac{1}{6}T^2D^2+\dots)^{-1}(1+\tfrac{1}{6}T^2D'^2+\dots)^{-1}u(x,y)$$

$$= u(x,y)-\tfrac{1}{6}T^2\Big(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\Big) \text{ approximately,}$$

if higher derivatives of $u(x,y)$ may be neglected.

2. The determination of a population from a given set of samples

On the Determination of Hypothetical Populations

The crucial problem with which this chapter has been concerned is how to make the fullest use of samples of a population for the drawing of conclusions regarding its structure; we are in fact trying to arrive at a mathematical method that will assist us in learning from experience. The general principle, already adopted in Chapter VIII, which we use for this purpose may be stated as follows:

(1) We assume a *class* of hypothetical populations capable of providing the samples found.

(2) This capacity involves the further assumption of a method of selecting the samples.

(3) We can then write down the probability that from any one of the class of populations, with this method of selection, precisely the given set of samples will be obtained. The probability will in general vary for different members of the class of populations; we then determine the member for which this probability is greatest, and choose it as the 'most likely' for the given set of samples in contrast to the 'most probable' sample.

The term 'most likely' is used here because now we are actually concerned with a new type of problem; we are not in fact discussing the question of the probability of occurrence of a particular population among a given class: the probability is now attached to the samples, not to the population. Thus, the 'most likely' population is defined as that member of a given class which yields the given samples with the greatest probability.

Suppose, for instance, that x_1, x_2,..., x_n are the measured observations of a number x; then the deviations of the observations from x are respectively $x-x_1$, $x-x_2,\ldots$, $x-x_n$.

If $p(\epsilon)$ is the probability that the method of observation gives a deviation of magnitude ϵ, the probabilities of obtaining the stated deviations are respectively

$$p(x-x_1),\quad p(x-x_2),\quad \ldots,\quad p(x-x_n).$$

Hence the probability of obtaining a combination of deviations $x-x_1$, $x-x_2,\ldots$, $x-x_n$ simultaneously is the product

$$p(x-x_1)p(x-x_2)\ldots p(x-x_n).$$

We propose to assume that the best approximation to x derivable from these samples is that value which makes the given combination of deviations the most probable that would occur in a sample of n observations. In effect we inquire, what process of selection applied to the readings x will give precisely this combination with the greatest probability? We are now in a position to apply this principle to a series of cases; we illustrate first with a case in which the samples have been obtained by a Bernoulli law of selection.

The Method of Maximum Likelihood

Suppose that the members of a population of given number N possess a certain characteristic in the unknown proportion $p:1$. If a series of samples n_1, $n_2,\ldots$ in number drawn from it is found to contain the characteristic in the proportions r_1/n_1, $r_2/n_2\ldots$, what information can be deduced with regard to the value of p? This is to raise a problem in induction if it is implied that the population has to be specified by means of the samples; and like all such problems it can be reduced to a deductive one by making an appropriate assumption. We postulate a class of population, capable of yielding the given samples, for which the probability of drawing the samples is calculable; we then inquire which member of this class will with the greatest probability furnish precisely the samples that have been found.

Thus, if a penny is tossed 100 times and found to give 50 heads, any probability p of obtaining a head, other than the value $p = \frac{1}{2}$, would give a smaller probability for the observed occurrence than if p were actually $\frac{1}{2}$.

If we apply these considerations to a class of Bernoulli distributions defined by the probability law ${}^nC_r p^r(1-p)^{n-r}$, where n is the size of the sample and r members possess the required quality, then the problem resolves itself into finding the value of p for which this probability is greatest. Since p does not occur in the coefficient ${}^nC_r$, we have simply to maximize $p^r(1-p)^{n-r}$, that is, to choose p so that $r/p = (n-r)/(1-p)$, or $r = np$.

The expression $p^r(1-p)^{n-r}$ is called by R. A. Fisher the 'likelihood': it is not in itself a probability, as we have seen, but an instrument for selecting the 'most likely' population from among a given class.

Ex. 1. An urn contains N black and white balls, pN of which are white. From it are drawn n_1 balls, each being replaced before another is drawn, and r_1 of these are found to be white. A second such extraction of n_2 balls is made, and among them are r_2 white balls. What is the most likely value of p?

The probability of obtaining the sample in question is

$${}^{n_1}C_{r_1}p^{r_1}(1-p)^{n_1-r_1}\times{}^{n_2}C_{r_2}p^{r_2}(1-p)^{n_2-r_2}.$$

Thus to find the value of p for which this probability is greatest we have to maximize the expression $p^{r_1+r_2}(1-p)^{n_1+n_2-r_1-r_2}$. In analogy with the preceding case this gives $r_1+r_2 = (n_1+n_2)p$.

Ex. 2. Suppose that the n_1 balls are marked as they are drawn and that in fact no ball is drawn twice. If these n_1 balls are now removed, the probability of obtaining a white ball at the second extraction is

$$p_1 = (pN-r_1)/(N--n_1),$$

and that of obtaining the given samples is proportional to

$$p^{r_1}(1-p)^{n_1-r_1}p_1^{r_2}(1-p_1)^{n_2-r_2},$$

that is, to

$$p^{r_1}(1-p)^{n_1-r_1}(pN-r_1)^{r_2}(N-n_1+r_1-pN)^{n_2-r_2}.$$

The value of p for which this is a maximum is given by the equation

$$\frac{r_1}{p}-\frac{n_1-r_1}{1-p}+\frac{Nr_2}{pN-r_1}-\frac{N(n_2-r_2)}{N-n_1+r_1-pN}=0.$$

It will be noticed that, if N is large, the value of p is given by

$$\frac{r_1}{p}-\frac{n_1-r_1}{1-p}+\frac{r_2}{p}-\frac{n_2-r_2}{1-p}=0,$$

so that $r_1+r_2 = (n_1+n_2)p$, as before.

Ex. 3. If $N = 14$, $n_1 = 5$, $r_1 = 2$, $n_2 = 2$, $r_2 = 1$, then in Ex. 1 we have $p = 3/7 = 0{\cdot}429$. In Ex. 2, the most likely value of p is found by maximizing the expression

$$p^2(1-p)^3(7p-1)(11-14p).$$

Thus p is the root of the cubic

$$343p^3-469p^2+164p-11 = 0$$

lying between $\frac{1}{2}$ and 1, i.e. $p = 0{\cdot}404$.

The method of maximum likelihood is applicable to hypothetical populations defined by more than one characteristic. For example, suppose that an urn contains balls of t different colours whose relative frequencies are $p_1, p_2,..., p_t$. If a sample of n balls is extracted and found to contain r_1 balls of the first type, r_2 of the second, and so on, we may inquire what values of $p_1, p_2,...$ make this sample the most probable. The probability of obtaining the sample is, by Bernoulli's Theorem,

$$P = \frac{n!}{r_1!\,r_2!\,...}p_1^{r_1}p_2^{r_2}...p_t^{r_t}, \tag{1}$$

where
$$p_1+p_2+...+p_t = 1 \tag{2}$$

and
$$r_1+r_2+...+r_t = n. \tag{3}$$

If P is a maximum, so is $\log P$; whence, if $\delta p_1, \delta p_2,...$ denote variations in $p_1, p_2...,$ we have the condition

$$\frac{r_1}{p_1}\delta p_1+\frac{r_2}{p_2}\delta p_2+...+\frac{r_t}{p_t}\delta p_t = 0, \tag{4}$$

where, by (2),
$$\delta p_1+\delta p_2+...+\delta p_t = 0. \tag{5}$$

Combining (4) and (5) we see that the conditions for a maximum are

$$\frac{r_1}{p_1} = \frac{r_2}{p_2} = ... = \frac{r_1+r_2+...+r_t}{p_1+p_2+...+p_t} = n, \text{ by (3)}.$$

It follows that

$$p_1 = r_1/n, \quad p_2 = r_2/n, \quad ..., \quad p_t = r_t/n.$$

Ex. An urn contains black, white, and yellow balls in unknown proportions $p_1 : p_2 : p_3$. Six balls are extracted, replaced, and six others

Black	*White*	*Yellow*
1	2	3
3	2	1

extracted. If the numbers of black, white, and yellow balls obtained in the two extractions are as shown, the values of p_1, p_2, p_3 for which the probability of obtaining this pair of samples is greatest are

$$p_1 = p_2 = p_3 = \tfrac{1}{3}.$$

The probability of drawing these samples will then be

$$P = \left[\frac{6!}{2!\,3!}\,\frac{1}{3^6}\right]^2.$$

Black	*White*	*Yellow*
2	2	2
2	2	2

If, instead, the two extractions had given rise to the second set of numbers shown, the values of p_1, p_2, p_3 obtained by maximizing the probability of obtaining the samples would have been as before, but the probability of drawing the samples would be

$$P' = \left[\frac{6!}{(2!)^3}\,\frac{1}{3^6}\right]^2;$$

and this is less than P. In fact we have

$$P'/P = 2^4/(3!)^2 = 4/9.$$

When we have determined the population for which a given sample is the most probable, it does not follow that even that sample is a very 'probable' one; its probability will depend on the number of types that might be drawn from such a hypothetical population, and on the relative frequency of occurrence of each type. Let us illustrate with a simple problem.

An urn contains black and white balls in an unknown proportion $p:1$. A certain number n is extracted, with replacement immediately after each extraction, and it is found that r of these are white. A second sample is obtained in the same manner. Let us suppose that in all 12 balls have been drawn and 6 of them found to be white; then the second sample consisted of $12-n$ balls, $6-r$ of which were white.

Since the ratio of the number of white balls extracted to the total number is $\frac{1}{2}$, it follows from the previous discussion (p. 166) that the 'most likely' value of p for the hypothetical population is $\frac{1}{2}$.

Consider now the probability of drawing just such a pair of samples from a population for which p is actually equal to $\frac{1}{2}$. The probability of drawing the first sample is ${}^nC_r(\frac{1}{2})^n$, and, since the balls are then returned to the urn, the probability of drawing the second is ${}^{12-n}C_{6-r}(\frac{1}{2})^{12-n}$. Hence the probability of obtaining the pair of extractions is

$$P = {}^{n}C_{r}\,{}^{12-n}C_{6-r}/2^{12},$$

and this will of course vary with n and r. It is not difficult to determine the values of n and r for which P is a maximum; since n and r can vary independently within the given limits, we require

$$ {}^{n-1}C_{r}\,{}^{13-n}C_{6-r} < {}^{n}C_{r}\,{}^{12-n}C_{6-r} > {}^{n+1}C_{r}\,{}^{11-n}C_{6-r}, $$

and $$ {}^{n}C_{r-1}\,{}^{12-n}C_{7-r} < {}^{n}C_{r}\,{}^{12-n}C_{6-r} > {}^{n}C_{r+1}\,{}^{12-n}C_{5-r}. $$

From these conditions it follows that

$$\frac{13r}{6}-1 < n < \frac{13r}{6}, \quad \text{and} \quad \frac{n}{2}-1 < r \leqslant \frac{n}{2}.$$

In virtue of the restrictions placed upon r, the second condition is a consequence of the first. We thus obtain the solution $r = 5$, $n = 10$, or $r = 1$, $n = 2$, and with either of these pairs of values $P = 504/2^{12} = P_0$, say.

In the accompanying table we give the proportions of white balls obtained in twelve pairs of extractions, with the corresponding values of P and P/P_0. Thus, although P_0 is itself small, it is 504 times as great as the probability of obtaining the first pair of extractions shown.

First drawing	*Second drawing*	$P \times 2^{12}$	P/P_0
6 : 6	0 : 6	1	0·002
5 : 5	1 : 7	7	0·014
4 : 4	2 : 8	28	0·056
5 : 6	1 : 6	36	0·072
4 : 5	2 : 7	105	0·21
1 : 4	5 : 8	224	0·448
2 : 6	4 : 6	225	0·45
2 : 3	4 : 9	378	0·756
3 : 6	3 : 6	400	0·8
2 : 4	4 : 8	420	0·84
0 : 1	6 : 11	462	0·924
1 : 2	5 : 10	504	1

The Method of Least Squares

The second law of selection to which we shall apply the foregoing principle is the Gaussian. It is worth while remarking that the method which we develop in part covers what is variously called curve fitting, smoothing of data, and graduation.

Each of these processes, whether it be the determination of a smooth curve that lies evenly among a set of points, or the smoothing out of an irregular curve, or the specification of an algebraic expression to cover a set of data, is in effect the determination of a hypothetical population, since each is merely a step towards specifying values of a variable at positions other than those immediately supplied by the data.

If the assumed hypothetical population is Gaussian, then in the notation of p. 165, $p(\epsilon) = \frac{h}{\sqrt{\pi}}\exp(-h^2\epsilon^2)$, so that the probability of obtaining the given sample is

$$\frac{h}{\sqrt{\pi}}\exp\{-h^2(x-x_1)^2\}\frac{h}{\sqrt{\pi}}\exp\{-h^2(x-x_2)\}\ldots\frac{h}{\sqrt{\pi}}\exp\{-h^2(x-x_n)^2\}$$

$$= \frac{h^n}{\pi^{n/2}}\exp\left\{-h^2\sum_{r=1}^{n}(x-x_r)^2\right\}.$$

For a given process of selection, h is a known constant; the problem, as before, is to find x so that the probability is a maximum. This is equivalent to determining x so that $\sum_{r=1}^{n}(x-x_r)^2$ is a minimum and, as we have seen, gives as the value of x the mean of x_1, x_2,..., x_n. This method of determining the best value of an observation by assuming that the sum of the squares of the deviations from it shall be a minimum is called the Method of Least Squares. Some writers prefer to begin with this method as the initial assumption, without directly implying the use of a Gaussian law.

Determination of the Precision Constant

The probability that the set of readings x_1, x_2,..., x_n will occur is

$$\frac{h^n}{\pi^{n/2}}\exp\{-h^2\sum(x_r-a)^2\},$$

where a is the mean of the readings.

Using the same principle as before, the value of h to be chosen is that which makes the above probability a maximum. Thus h is determined by the equation

$$\frac{d}{dh}\{h^n\exp[-h^2\sum(x_r-a)^2]\} = 0,$$

so that $$\frac{n}{h} = 2h \sum (x_r-a)^2.$$

Hence $$h^2 = \frac{n}{2\sum (x_r-a)^2} = \frac{1}{2\sigma'^2},$$

where σ' is the standard deviation for the given set of readings. It follows that our choice of h is such as to make the standard deviation for the set x_1, x_2,..., x_n coincide with that of the assumed hypothetical population.

Curve Fitting

Now suppose that $Y = f(x, a)$ represents a possible series of hypothetical populations, obtained by varying a, from one of which the given sample is presumed to have been drawn. As before, we shall assume that the probability of committing an error of magnitude ϵ is $\frac{h}{\sqrt{\pi}}\exp(-h^2\epsilon^2)$, and that the precision constant h is the same for each measurement irrespective of its position in the range. Suppose that readings y_1, y_2,..., y_n are taken at the positions† x_1, x_2,..., x_n, and that Y_1, Y_2,..., Y_n are the corresponding values of the hypothetical population. This assumes that the x's are accurate. Then the probability of drawing this sample from the population whose parameter is a is

$$\frac{h}{\sqrt{\pi}}\exp\{-h^2(Y_1-y_1)^2\}\frac{h}{\sqrt{\pi}}\exp\{-h^2(Y_2-y_2)^2\}...$$
$$= \frac{h^n}{\pi^{n/2}}\exp\left\{-h^2\sum_1^n (Y_r-y_r)^2\right\},$$

where Y_1, Y_2,..., Y_r depend on the parameter a.

We propose to choose as *the* hypothetical population among the set $f(x, a)$ the one that makes the occurrence of this set of readings the most probable. We have thus to make $\sum (Y_r-y_r)^2$ a minimum, i.e. we have to choose a so that $\sum [f(x, a)-y_r]^2$ is a minimum. Hence a must satisfy the equation

$$\frac{\partial}{\partial a}\sum_1^r [f(x_r, a)-y_r]^2 = 0,$$

and thus, on the foregoing assumptions, the hypothetical population is determined.

† If the readings are weighted, i.e. if several readings occur at the same position, the x's are not all different.

Ex. 1.

x	1	2	3	4	5
y	$-0\cdot8$	$0\cdot9$	$3\cdot1$	$5\cdot3$	$6\cdot8$

The values of y shown are subject to accidental errors. Given that the population Y from which they are extracted is one of the system $Y = 2x+a$, determine the best value of a.

We have to choose a so that the expression

$$(2\cdot8+a)^2+(3\cdot1+a)^2+(2\cdot9+a)^2+(2\cdot7+a)^2+(3\cdot2+a)^2$$

is a minimum, whence $a = -\dfrac{14\cdot7}{5} = -2\cdot94$.

Ex. 2. Find the best values of a and m if the values Y are given by the function $Y = mx+a$.

Ex. 3. If it is desired to represent the following values of y approximately by a function of the form $y = a+bx+cx^2$, determine the best values of a, b, and c.

x	0	1	2	3	4	5	6	7	8	9	10
y	$7\cdot98$	$11\cdot51$	$14\cdot02$	$15\cdot46$	$16\cdot01$	$15\cdot51$	$13\cdot98$	$11\cdot52$	$8\cdot02$	$3\cdot31$	-2

Here a, b, and c have to be chosen to make the sum of the squares of the deviation from y a minimum.

NOTE. Suppose that we wish to fit a Gaussian law of the form $y = \exp(a+bx+cx^2)$ to a distribution curve. We might proceed by taking logarithms and determining the best values of a, b, c (as in the above example) for the readings. Such a method, although convenient in practice, is not strictly justifiable, since, if the errors in y are distributed according to a Gaussian law, those of $\log y$ are not.

Ex. 4. Find the values of a and b for which the parent population $y = ax+b\sin x$ would give the pairs of values

x	$0\cdot2$	$0\cdot8$	$1\cdot4$	$2\cdot0$
y	$0\cdot202$	$0\cdot882$	$1\cdot821$	$3\cdot421$

as the most probable, assuming that the deviations follow the Gaussian law.

The Line of Regression

Suppose that

$$x_1, x_2,..., x_n,$$
$$y_1, y_2,..., y_n$$

are n pairs of data related in the sense that changes in the values of the x's are accompanied by changes in the y's. Assuming that there are no errors in the x's, we wish to determine to what extent the numbers (x, y) may be considered as derivable from the hypothetical population

$$y = Ax+B,$$

assuming that the deviations follow the Gaussian law.

We have thus to minimize the expression

$$\sum_{r=1}^{n} (Ax_r+B-y_r)^2.$$

This means that A and B must satisfy the equations

$$\sum x_r(Ax_r+B-y_r) = 0,$$

$$\sum (Ax_r+B-y_r) = 0,$$

that is,

$$A\sum x_r^2+B\sum x_r = \sum x_r y_r, \qquad (1)$$

$$A\sum x_r+nB = \sum y_r. \qquad (2)$$

It is convenient to replace x and y by their deviations from the corresponding means X, Y; writing $x = X+\xi$, $y = Y+\eta$, we have

$$\sum x_r = nX, \qquad \sum y_r = nY,$$

$$\sum x_r^2 = \sum (X+\xi_r)^2 = nX^2+\sum \xi_r^2,$$

and $$\sum x_r y_r = \sum (X+\xi_r)(Y+\eta_r) = nXY+\sum \xi_r \eta_r.$$

Thus (1) and (2) become

$$A(X^2+\sigma_x^2)+BX = XY+\frac{1}{n}\sum \xi_r \eta_r, \qquad (3)$$

$$AX+B = Y, \qquad (4)$$

where σ_x is the standard deviation of the x's from X.

Solving (3) and (4) for A and B we obtain

$$A = \frac{1}{n\sigma_x^2}\sum \xi_r \eta_r = \frac{\sum \xi_r \eta_r}{\sum \xi_r^2},$$

$$B = Y-\frac{X\sum \xi_r \eta_r}{\sum \xi_r^2}.$$

Hence the hypothetical population is given by the curve

$$y-Y = \frac{\sum \xi_r \eta_r}{\sum \xi_r^2}(x-X).$$

This curve is called the 'line of regression' for the given data, and can be written as

$$\frac{y-Y}{\sigma_y} = r\frac{x-X}{\sigma_x}, \qquad (5)$$

where σ_x and σ_y are the standard deviations of the x's and y's from their respective means and

$$r = \frac{\sum \xi_r \eta_r}{\sqrt{(\sum \xi_r^2 \sum \eta_r^2)}}.$$

Correlation

We are now in a position to examine the problem of 'correlation'. Suppose that the two sets of data as in the preceding paragraph have been found. If the points (x_r, y_r) are plotted on a diagram and form a 'good' curve or lie very nearly on a straight line, then the principles of curve fitting for the selection of a hypothetical population can be applied at once because the necessary specification of the population is not difficult to make. When, however, the points (x_r, y_r) are so scattered as to render this impossible, we are at liberty to make any reasonable assumption in terms of which to interpret the data.

Two methods of procedure are usually adopted. We begin with the assumption that the x's and y's are attempted measures of points on some straight line but that the x's are measured without error. Then it follows from the previous section that the hypothetical population is given by

$$\frac{y-Y}{\sigma_y} = r\frac{x-X}{\sigma_x},$$

where X, Y are the means of the x's and y's, σ_x, σ_y are the corresponding standard deviations, and

$$r = \frac{\sum (X-x_r)(Y-y_r)}{\sqrt{\{\sum (X-x_r)^2 \sum (Y-y_r)^2\}}}.$$

The curve so obtained represents a special member of an assumed class of hypothetical population, called the 'line of regression' of y on x, for it measures the extent to which a variation in x effects a change in y; in fact, when x changes by σ_x, y/σ_y changes by r.

We could, however, have approached the same problem by choosing a second class of hypothetical population on the assumption that the y's were correct values and that the x's involved errors. It is easy to see that the member of the population then selected would be

$$\frac{x-X}{\sigma_x} = r\frac{y-Y}{\sigma_y}.$$

This represents the line of regression of x on y; when y changes by σ_y, x/σ_x changes by r. Thus r is a measure common to both the hypothetical populations; it is called the 'coefficient of linear correlation' and is taken to be a measure of the extent to which the sets of numbers x and y are interlinked.

It is clear that if the two lines of regression are coincident, then $r^2 = 1$, and if they are at right angles, $r = 0$. In the case $r = 1$, there is maximum correlation and x and y are linearly related over the whole range. When $r = 0$ the variation in x has no influence on the variation in y. Thus r is a number whose absolute magnitude lies between 0 and 1. We note that r may, however, be negative, in which case an increase in x is accompanied by a decrease in y, and vice versa.

Ex. Two sets of numbers are chosen in the intervals (0, 4), (5, 9), (10, 14),..., (30, 34), with the following results:

x	1	6	12	16	20	25	32
y	3	6	13	16	22	28	31

We thus obtain

$$X = 16, \qquad Y = 17, \qquad \sum (X-x)(Y-y) = 679,$$
$$\sum (X-x)^2 = 694, \qquad \sum (Y-y)^2 = 676,$$

so that r is given by

$$r = \frac{679}{\sqrt{(694 \times 676)}} = \frac{679}{684} = 0\cdot 99, \quad \text{approximately.}$$

Generally, if we have two sets of numbers $x_1, x_2,..., x_n$ and $y_1, y_2,..., y_n$ such that x_r and y_r lie in the interval (t_r, t_{r+1}), and if the differences $t_r - t_{r+1}$ are small and equal for all values of r, then x_r and y_r will correlate almost exactly linearly.

The method we have used to find the coefficient of linear correlation is capable of immediate extension. Thus, for parabolic correlation, we wish to find the value of λ for which the hypothetical population

$$Y' = \lambda X'^2, \quad \text{where} \quad X' = \frac{X-x}{\sigma_x}, \qquad Y' = \frac{Y-y}{\sigma_y},$$

best fits the given numbers $(X_1', Y_1'),..., (X_n', Y_n')$.

We have therefore to choose λ so that $\sum (Y_r' - \lambda X_r'^2)^2$ is a minimum.

Hence

$$\lambda = \frac{\sum X_r'^2 Y_r'}{\sum X_r'^4} = \frac{\sigma_x^2 \sum (X-x_r)^2 (Y-y_r)}{\sigma_y \sum (X-x_r)^4}$$
$$= \frac{\sum (X-x_r)^2 \sum (X-x_r)^2 (Y-y_r)}{\sqrt{n}\sqrt{\{\sum (Y-y_r)^2\}} \sum (X-x_r)^4}.$$

The interpretation of λ in this case is, of course, quite different from that for r in the previous case.

The Method of Maximum Correlation

Let $(x_1, x_2,..., x_n)$ and $(y_1, y_2,..., y_n)$ be two sets of observations obtained by two different experimenters to represent variations in the same phenomena at positions x which are accurately given. If the experiments havė been carefully performed and the differences between corresponding pairs (x_s, y_s) of observations are due only to accidental errors, it follows that the two sets will be highly correlated; in other words, the line of regression of x on y or of y on x will be very near the line $y = x$. We wish to determine from these observations a third set $(z_1, z_2,..., z_n)$ which *correlates most highly* with the given sets; that is to say, if r_{xz} and r_{yz} are the correlation coefficients between the x's and the z's and between the y's and the z's, respectively, then the z's are to be chosen so as to make some symmetric function $F(r_{xz}, r_{yz})$ a maximum. Each such function defines a class of populations. Consider in particular

$$F \equiv r_{xz}+r_{yz}.$$

Let X, Y, and Z be the means of the three sets of observations, and ξ_s, η_s, and ζ_s the deviations of x_s, y_s, and z_s each from its mean; then

$$\sum_1^n \xi_s = \sum_1^n \eta_s = \sum_1^n \zeta_s = 0,$$

$$r_{xy} = \frac{\sum \xi_s \eta_s}{\sqrt{(\sum \xi_s^2 \sum \eta_s^2)}}, \quad r_{xz} = \frac{\sum \xi_s \zeta_s}{\sqrt{(\sum \xi_s^2 \sum \zeta_s^2)}}, \quad r_{yz} = \frac{\sum \eta_s \zeta_s}{\sqrt{(\sum \eta_s^2 \sum \zeta_s^2)}}.$$

To simplify the notation we write

$$\frac{\xi_s}{\sqrt{\sum \xi_s^2}} = a_s, \qquad \frac{\eta_s}{\sqrt{\sum \eta_s^2}} = b_s, \qquad \frac{\zeta_s}{\sqrt{\sum \zeta_s^2}} = c_s.$$

Then the above relations may be replaced by

$$\sum a_s = \sum b_s = \sum c_s = 0,$$
$$\sum a_s^2 = \sum b_s^2 = \sum c_s^2 = 1,$$
$$r_{xy} = \sum a_s b_s, \qquad r_{xz} = \sum a_s c_s, \qquad r_{yz} = \sum b_s c_s.$$

Now if the function F is to be a maximum we require

$$\delta F = \delta r_{xz}+\delta r_{yz} = 0, \tag{1}$$

where $\quad \delta r_{xz} = \sum a_s \delta c_s \quad$ and $\quad \delta r_{yz} = \sum b_s \delta c_s. \qquad (2)$

Substituting from (2) in (1) we thus require

$$\sum (a_s+b_s)\delta c_s = 0. \tag{3}$$

From the restrictive conditions on c_s we have

$$\sum c_s \delta c_s = 0 \quad \text{and} \quad \sum \delta c_s = 0. \tag{4}$$

Hence, from (3) and (4) we obtain

$$\sum (a_s+b_s+\lambda c_s+\mu)\, \delta c_s = 0,$$

where λ and μ are constants to be determined.

Equating the coefficients of δc_s to zero we have

$$a_s+b_s+\lambda c_s+\mu = 0 \quad (s = 1, 2,..., n). \tag{5}$$

Summing the s relations (5) we obtain

$$\sum a_s+ \sum b_s+\lambda \sum c_s+s\mu = 0,$$

whence we deduce that $\mu = 0$.

Multiplying (5) by c_s and summing, we have

$$r_{xz}+r_{yz}+\lambda = 0.$$

Accordingly (5) takes the form

$$a_s+b_s = (r_{xz}+r_{yz})c_s \quad (s = 1, 2,..., n). \tag{6}$$

Multiplying (6) by a_s and b_s respectively and summing, we obtain

$$\left.\begin{aligned} 1+r_{xy} &= (r_{xz}+r_{yz})r_{xz} \\ 1+r_{xy} &= (r_{xz}+r_{yz})r_{yz}. \end{aligned}\right\} \tag{7}$$

From (7) it follows that $\quad r_{xz} = r_{yz}.$ (8)

Equations (6), (7), and (8) serve to determine r_{xz}, r_{yz}, and c_s. Thus from (7) and (8) we have

$$r_{xz} = r_{yz} = \sqrt{\{\tfrac{1}{2}(1+r_{xy})\}}, \tag{9}$$

and from (6) $\quad c_s = (a_s+b_s)/\sqrt{\{2(1+r_{xy})\}}.$ (10)

Since r_{xy} is positive in the case with which we are concerned, and is moreover less than unity, it follows from (10) that c_s is slightly greater than the mean of a_s and b_s.

Returning now to our original notation we have still to determine z_s.

We have $\quad z_s = Z+\zeta_s = Z+\gamma c_s,$ (11)

where $\gamma = \sqrt{(\sum \zeta_s^2)}$ and Z are unknown.

We propose to determine the latter by the method of least squares. We have thus to make

$$\sum \{(z_s-x_s)^2+(z_s-y_s)^2\},$$

i.e. $$\sum \{(Z+\gamma c_s-X-\xi_s)^2+(Z+\gamma c_s-Y-\eta_s)^2\},$$

a minimum for variations in Z and γ.

We thus obtain the equations

$$Z = \tfrac{1}{2}(X+Y) \tag{12}$$

and $$2\gamma \sum c_s^2 = \sum \xi_s c_s + \sum \eta_s c_s,$$

or $$2\gamma = r_{xz}\{\sqrt{(\sum \xi_s^2)}+\sqrt{(\sum \eta_s^2)}\}, \quad \text{by (8).}$$

In terms of the standard deviations this result may be written as

$$Y = \tfrac{1}{2}\sqrt{n}r_{xz}(\sigma_x+\sigma_y). \tag{13}$$

Hence

$$\begin{aligned} Z_s &= Z+\gamma c_s \\ &= \tfrac{1}{2}(X+Y)+\tfrac{1}{2}\sqrt{n}r_{xz}(\sigma_x+\sigma_y)\frac{(a_s+b_s)}{\sqrt{\{2(1+r_{xy})\}}} \\ &= \tfrac{1}{2}(X+Y)+\tfrac{1}{4}\sqrt{n}(\sigma_x+\sigma_y)\left[\frac{\xi_s}{\sqrt{(\sum \xi_s^2)}}+\frac{\eta_s}{\sqrt{(\sum \eta_s^2)}}\right] \\ &= \tfrac{1}{2}(X+Y)+\tfrac{1}{4}(\sigma_x+\sigma_y)(\xi_s/\sigma_x+\eta_s/\sigma_y). \end{aligned} \tag{14}$$

Thus all the constants in the calculation of the set (Z_s) have been determined.

For the application of this method to the general case of m given sets of observations and for more general forms of the function F, reference may be made to a recent paper.†

Linear Correlation in General

Suppose that

$$x_1, x_2,..., x_n,$$
$$y_1, y_2,..., y_n$$

is a given system of data. We may inquire which member of the class of hypothetical populations $x\cos\alpha+y\sin\alpha = p$, where α and p are variable, will provide this system of data with the greatest probability.

Let us suppose that (X_r, Y_r) is the point on this line to which (x_r, y_r) is an empirical approximation, and that the errors in the placing of x_r and y_r occur independently with frequencies determined by the same Gaussian law. Thus the probability of an error X_r-x_r is proportional to $\exp\{-h^2(X_r-x_r)^2\}$ and that of an error Y_r-y_r is proportional to $\exp\{-h^2(Y_r-y_r)^2\}$. The proba-

† H. Levy and J. C. Gascoigne, *Proc. Phys. Soc.* **48** (1935).

bility of obtaining the whole set of data is therefore proportional to

$$\exp\{-h^2(X_1-x_1)^2\}\exp\{-h^2(Y_1-y_1)^2\}\dots\exp\{-h^2(X_n-x_n)^2\}\times$$
$$\times\exp\{-h^2(Y_n-y_n)^2\} = \exp\Big\{-h^2\sum_1^n[(X_r-x_r)^2+(Y_r-y_r)^2]\Big\}.$$

If this is to be a maximum we require that

$$\sum[(X_r-x_r)^2+(Y_r-y_r)^2]$$

shall be a minimum.

Geometrically, this expression represents the sum of the squares of the distances of the points (x_r, y_r) from the corresponding points (X_r, Y_r) on the line

$$x\cos\alpha+y\sin\alpha = p. \tag{1}$$

Now unless (X_r, Y_r) is the foot of the perpendicular from (x_r, y_r) on (1), the given expression will certainly not attain its least value. Since the perpendicular from (x_r, y_r) on (1) is of length $x_r\cos\alpha+y_r\sin\alpha-p$, we have to determine α and p so that

$$\sum(x_r\cos\alpha+y_r\sin\alpha-p)^2$$

is a minimum. We thus require

$$\sum(x_r\cos\alpha+y_r\sin\alpha-p) = 0 \tag{2}$$

and $$\sum(x_r\cos\alpha+y_r\sin\alpha-p)(x_r\sin\alpha-y_r\cos\alpha) = 0. \tag{3}$$

Equation (2) may be written in the form

$$\frac{\sum x_r}{n}\cos\alpha+\frac{\sum y_r}{n}\sin\alpha-p = 0,$$

which shows that the mean position (X, Y) of the points (x_r, y_r) lies on (1). Writing, as before, $x_r = X+\xi_r$, $y_r = Y+\eta_r$, so that $\sum\xi_r = \sum\eta_r = 0$, equations (2) and (3) become

$$X\cos\alpha+Y\sin\alpha = p, \tag{4}$$

$$\sum(\xi_r\cos\alpha+\eta_r\sin\alpha)(\xi_r\sin\alpha-\eta_r\cos\alpha+X\sin\alpha-Y\cos\alpha) = 0,$$

or $$(\cos^2\alpha-\sin^2\alpha)\sum\xi_r\eta_r = \sin\alpha\cos\alpha\sum(\xi_r^2-\eta_r^2), \tag{5}$$

whence $$\tan 2\alpha = \frac{2\sum\xi_r\eta_r}{\sum(\xi_r^2-\eta_r^2)}. \tag{6}$$

Thus the required hypothetical population is given by

$$(x-X)\cos\alpha+(y-Y)\sin\alpha = 0, \tag{7}$$

where α is determined by (6).

Eliminating α between (6) and (7), we find that (7) is one of the pair of lines whose equation is

$$\frac{(x-X)(y-Y)}{(x-X)^2-(y-Y)^2} = \frac{\sum \xi_r \eta_r}{\sum (\xi_r^2-\eta_r^2)}. \tag{8}$$

It will be observed that these lines are the bisectors of the angles between the regression lines of x on y, and of y on x; one of them determines the most probable and the other the least probable of hypothetical populations represented by lines passing through (X, Y). The required line is the bisector of the acute angle between the regression lines.

When $\sigma_x = \sigma_y$, it follows from (8) that the line (7) has the equation

$$y-Y = \pm(x-X).$$

The Gaussian Law for Two Variables: Correlation

We can approach the problem of correlation in the following way.

Let η_1, η_2 be the deviations of two sets of quantities from their respective means, and suppose η_1 and η_2 are each determined from elements ϵ_1, ϵ_2 themselves also deviations from their means, and distributed about these means according to the Gaussian law. Suppose

$$\begin{aligned}\eta_1 &= a\epsilon_1+b\epsilon_2 \\ \eta_2 &= \alpha\epsilon_1+\beta\epsilon_2\end{aligned}, \quad \text{or} \quad \begin{aligned}\epsilon_1 &= A\eta_1+B\eta_2 \\ \epsilon_2 &= R\eta_1+S\eta_2.\end{aligned}$$

Then the probability of the occurrence of the ϵ's simultaneously within the ranges $(\epsilon_1, \epsilon_1+\delta\epsilon_1)$ and $(\epsilon_2, \epsilon_2+\delta\epsilon_2)$ is

$$\frac{h_1}{\sqrt{\pi}}\exp(-h_1^2\epsilon_1^2)\,d\epsilon_1\,\frac{h_2}{\sqrt{\pi}}\exp(-h_2^2\epsilon_2^2)\,d\epsilon_2$$
$$= \frac{h_1 h_2}{\pi}\exp(-h_1^2\epsilon_1^2-h_2^2\epsilon_2^2)\,d\epsilon_1\,d\epsilon_2.$$

If for the ϵ's we substitute their values in terms of the η's as above, the result will give the probability of the occurrence of the two characteristics η_1 and η_2 in the ranges $(\eta_1, \eta_1+\delta\eta_1)$, $(\eta_2, \eta_2+\delta\eta_2)$, viz.

$$\exp\{-(\lambda\eta_1^2+2\mu\eta_1\,\eta_2+\nu\eta_2^2)\}\,\delta\eta_1\,\delta\eta_2,$$

an extension of the Gaussian law.

We are now in a position to generalize and interpret this expression.

As before, let η_1 and η_2 be the deviations of two measurable characteristics each from its mean. The problem is to represent the linkage that shows itself, if at all, between the quantities η_1 and η_2. Let them be determined by the corresponding deviations of a number of contributory elements, $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ each from its mean. Then

$$\eta_1 = a_1 \epsilon_1 + a_2 \epsilon_2 + \ldots + a_m \epsilon_m,$$

$$\eta_2 = b_1 \epsilon_1 + b_2 \epsilon_2 + \ldots + b_m \epsilon_m.$$

Let us assume that each of the ϵ_r contributions conforms to the Gaussian law of error with precision constant h_r; then the compound probability that the ϵ's lie respectively and simultaneously in the ranges

$$(\epsilon_1, \epsilon_1 + \delta\epsilon_1),\ (\epsilon_2, \epsilon_2 + \delta\epsilon_2),\ \ldots,\ (\epsilon_m, \epsilon_m + \delta\epsilon_m),$$

is $$R = A \exp(-\sum \epsilon_r^2 h_r^2)\, d\epsilon_1\, d\epsilon_2 \ldots d\epsilon_m,$$

since they vary independently.

In this expression substitute for ϵ_1 and ϵ_2 in terms of η_1 and η_2; then the compound probability that η_1 and η_2 lie in the range $(\eta_1, \eta_1 + \delta\eta_1)$ and $(\eta_2, \eta_2 + \delta\eta_2)$, respectively, as well as that the remaining ϵ's should lie in the ranges

$$(\epsilon_3, \epsilon_3 + \delta\epsilon_3),\ \ldots,\ (\epsilon_m, \epsilon_m + \delta\epsilon_m),$$

is of the form

$$Q = B e^{-U} \delta\eta_1 \delta\eta_2 \delta\epsilon_3 \ldots \delta\epsilon_m,$$

where U is the sum of

(i) a quadratic function of η_1 and η_2

(ii) a quadratic function of $\epsilon_3, \ldots, \epsilon_m$

(iii) a function linear in η_1, η_2, $\epsilon_3, \ldots, \epsilon_m$.

If Q be integrated with respect to the ϵ's from $+\infty$ to $-\infty$ the result will give the probability of the occurrence of the two characteristics η_1 and η_2 in the ranges $(\eta_1, \eta_1 + \delta\eta_1)$, $(\eta_2, \eta_2 + \delta\eta_2)$. Finally we obtain

$$P = C \exp\{-\tfrac{1}{2}(C_1 \eta_1^2 + 2C_{12} \eta_1 \eta_2 + C_2 \eta_2^2)\} \delta\eta_1 \delta\eta_2,$$

clearly an extension of the Gaussian law of error. If η_1 and η_2 were quantities that could be chosen independently with standard deviations σ_1 and σ_2 respectively, then we should have as the probability of the occurrence of η_1 and η_2 in the given range

$$P' = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}-\frac{\eta_2^2}{2\sigma_2^2}\right).$$

The presence of the term $\eta_1\eta_2$ in the quadratic expression brings out the linkage between the two quantities.

Write

$$C_1\eta_1^2+2C_{12}\eta_1\eta_2+C_2\eta_2^2 = C_1\left[\eta_1^2+\frac{2C_{12}}{C_1}\eta_1\eta_2\right]+C_2\eta_2^2$$
$$= C_1\left[\eta_1+\frac{C_{12}}{C_1}\eta_2\right]^2+\left[C_2-\frac{C_{12}^2}{C_1}\right]\eta_2^2$$
$$= C_2\left[\eta_2+\frac{C_{12}}{C_2}\eta_1\right]^2+\left[C_1-\frac{C_{12}^2}{C_2}\right]\eta_1^2.$$

Now write $\qquad C_1-\dfrac{C_{12}^2}{C_2} = \dfrac{1}{\sigma_1^2}, \qquad C_2-\dfrac{C_{12}^2}{C_1} = \dfrac{1}{\sigma_2^2},$

and $\qquad r = -\dfrac{C_{12}}{\sqrt{(C_1C_2)}}.$

Thus $\qquad \dfrac{1}{\sigma_1^2} = C_1(1-r^2); \qquad \dfrac{1}{\sigma_2^2} = C_2(1-r^2).$

Moreover, since P is to be a probability, when integrated with respect to η_1 and η_2 from $+\infty$ to $-\infty$, it has the value unity.

Hence

$$1 = C\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{-U}\,d\eta_1\,d\eta_2.$$

This integral is evaluated below with the result

$$1 = \frac{2\pi C}{\sqrt{(C_1C_2-C_{12}^2)}} = 2\pi C\sigma_1\sigma_2\sqrt{(1-r^2)}.$$

Accordingly the law of error takes the form†

$$\frac{1}{2\pi\sigma_1\sigma_2}\,\frac{1}{(1-r^2)^{\frac12}}\exp\left\{-\frac{1}{2(1-r^2)}\left(\frac{\eta_1^2}{\sigma_1^2}-\frac{2r\eta_1\eta_2}{\sigma_1\sigma_2}+\frac{\eta_2^2}{\sigma_2^2}\right)\right\}.$$

† In this connexion we may note Mehler's series for the correlation function

$$(1-r^2)^{-\frac12}\exp\left(-\frac{1}{2}\,\frac{(x^2-2rxy+y^2)}{1-r^2}\right)$$
$$= \exp\{-\tfrac12(x^2+y^2)\}\left\{1+rH_1(x)H_1(y)+\frac{r^2}{2!}H_2(x)H_2(y)+\dots\right\},$$

where $H_r(x)$ is the Hermite polynomial (see p. 138).

It is clear that when there is no correlation ($r = 0$), i.e. when the term $\eta_1 \eta_2$ is absent, σ_1 and σ_2 become simply the standard deviations of the η_1's and η_2's in that case. That they still bear this interpretation even in the case of linkage may be seen from the following considerations.

Write
$$z = C\exp\{-\tfrac{1}{2}(C_1 x^2+2C_{12}xy+C_2 y^2)\}.$$

Then from the integrals evaluated below we have for the second moment of x

$$I_1 = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} zx^2\,dxdy = \frac{2\pi C_2 C}{(C_1 C_2-C_{12}^2)^{\frac{3}{2}}} = \sigma_1^2,$$

$$I_2 = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} zy^2\,dxdy = \frac{2\pi C_1 C}{(C_1 C_2-C_{12}^2)^{\frac{3}{2}}} = \sigma_2^2.$$

Similarly,

$$J = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} zxy\,dxdy = \frac{-2\pi C_{12}}{(C_1 C_2-C_{12}^2)^{\frac{3}{2}}} = r\sigma_1\sigma_2.$$

The integrals I_1 and I_2 are the squares of the standard deviations of the η's while J is the sum of the products $\eta_1 \eta_2$.

Accordingly, $$r = \frac{J}{\sigma_1 \sigma_2} = \frac{J}{\sqrt{(I_1 I_2)}},$$

or for computational purposes we write

$$r = \frac{\sum \eta_1 \eta_2}{\sqrt{(\sum \eta_1^2 \sum \eta_2^2)}}.$$

It remains to evaluate the integrals referred to above.

Consider

$$A = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}(ax^2-2hxy+by^2)\}\,dxdy.$$

Now $$ax^2-2hxy+by^2 = a\left(x-\frac{h}{a}y\right)^2+\frac{\Delta}{a}y^2,$$

where $$\Delta = ab-h^2.$$

Also $$\int\limits_{-\infty}^{\infty} \exp(-a^2x^2)\,dx = \frac{\sqrt{\pi}}{a}.$$

Thus

$$\int\limits_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}(ax^2-2hxy+by^2)\}\,dx = \sqrt{\frac{2\pi}{a}}\exp\left(-\frac{1}{2}\frac{\Delta}{a}y^2\right).$$

Hence

$$A = \sqrt{\frac{2\pi}{a}} \int\limits_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{\Delta}{a}y^2\right) dy$$

$$= \sqrt{\frac{2\pi}{a}} \frac{\surd\pi}{\sqrt{\frac{\Delta}{2a}}} = \frac{2\pi}{\surd\Delta} = \frac{2\pi}{\sqrt{(ab-h^2)}}.$$

By differentiating under the integral sign, with respect to a, h, and b, we obtain

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x^2 \exp\{-\tfrac{1}{2}(ax^2-2hxy+by^2)\}\, dxdy = \frac{2\pi b}{(ab-h^2)^{\frac{3}{2}}}$$

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} xy \exp\{-\tfrac{1}{2}(ax^2-2hxy+by^2)\}\, dxdy = \frac{2\pi h}{(ab-h^2)^{\frac{3}{2}}}$$

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} y^2 \exp\{-\tfrac{1}{2}(ax^2-2hxy+by^2)\}\, dxdy = \frac{2\pi a}{(ab-h^2)^{\frac{3}{2}}}.$$

Tests of Significance for Small Samples

One of the most important contributions which statistical analysis has made to experimental practice lies in what are called 'tests of significance'. Suppose that a series of measurements is made of a quantity which in 'normal' circumstances would have the value m. From a study of the observations, can we say that these are themselves normal measures of m? If not, can some measure be found to estimate the degree of non-normality? For example, a collection of trees is sprayed with an insecticide, and after a lapse of time the number of insects upon them is counted. A corresponding series of unsprayed trees (controls), of equal number, is also counted for insects. We may ask whether the difference between the average number of insects per tree in the two series is sufficiently great for us to assert that the effect of spraying has been significant. From the point of view of probability we may regard the problem in this light: we may say that there are n numbers $x_1, x_2,..., x_n$, whose average is $\bar{x}$; m is the mean to be anticipated if n were of infinite extent and if no factor had operated to disturb the equilibrium of the series. In asking, therefore, what is the significance of $\bar{x}-m$, we are really inquiring with what probability one might expect a deviation of $\bar{x}$ from m to have as large a magnitude as this, under so-called 'random' conditions, i.e. when the numbers x_1, x_2,..., x_n are chosen about m according to a 'random' law—for present purposes the Gaussian.

Once that probability has been found, it will be possible to express in any given case the significance of the deviation in terms of the probability that it will arise at random. If it is very probable that a deviation of this amount will occur in a random sample of n, then there is little experimental significance in the deviation found; and conversely. It should be remarked that, in expressing the significance of the deviation in this way in terms of probability, we are really referring it to the significance of the probability—a matter which, as we have previously remarked, is to be decided finally by the experimenter himself. It is clear that a corresponding investigation of significance can be made for the probability of occurrence of a deviation in any other typical constant from that of an assumed infinite population.

Let there be a population, in number N, distributed according to the normal law

$$y = \frac{N}{\sigma\sqrt{(2\pi)}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{1}$$

where σ is the standard deviation of the population and m is its mean.

Suppose that a sample n in number is drawn from it, having magnitudes x_1, x_2,..., x_n. We can write down the probability that the members of the sample should lie between x_1 and x_1+dx_1, x_2 and x_2+dx_2,..., x_n and x_n+dx_n. This is

$$\mathbf{P} = \frac{N}{\sigma\sqrt{(2\pi)}}\exp\left(-\frac{(x_1-m)^2}{2\sigma^2}\right)\frac{N}{\sigma\sqrt{(2\pi)}}\exp\left(-\frac{(x_2-m)^2}{2\sigma^2}\right)...\times$$

$$\times\frac{N}{\sigma\sqrt{(2\pi)}}\exp\left(-\frac{(x_n-m)^2}{2\sigma^2}\right)dx_1\,dx_2...dx_n,$$

i.e. $$\mathbf{P} = \frac{N^n}{\{\sigma\sqrt{(2\pi)}\}^n}\exp\left(-\frac{1}{2\sigma^2}\sum(x_r-m)^2\right)dx_1\,dx_2...dx_n.$$

Thus

$$\mathbf{P} = A\exp\left(-\frac{1}{2\sigma^2}[\sum(x_r-\bar{x})^2+n(\bar{x}-m)^2]\right)dx_1\,dx_2...dx_n, \tag{2}$$

where A is a constant and $\bar{x} = \sum x_r/n$.

We now represent the sample by a point P in space of n dimensions having coordinates $(x_1, x_2,..., x_n)$. Then

$$x_1 = x_2 = ... = x_n$$

is the line which is equally inclined to the coordinate axes. The perpendicular distance of P from this line is given by

$$PM^2 = (x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2,$$

where M is the point $(\bar{x}, \bar{x},..., \bar{x})$.

Thus $PM = s\sqrt{n}$, where s is the standard deviation of the sample. Hence, given $\bar{x}$ and, therefore, M, for a fixed s, the point P must lie on a sphere of $(n-1)$ dimensions with centre at M and radius $s\sqrt{n}$.

An element of volume in this space may thus be expressed in terms of the variation of $\bar{x}$, namely $d\bar{x}$, and the variation $d(s^{n-1})$ in surface area. Thus the volume element can be written as

$$Cs^{n-2}\,ds d\bar{x},$$

where C is some constant.

We now see that this representation of our sample, together with the symmetrical nature of the expressions for x and s, enables us to replace (2) by the formula

$$\mathbf{P} = Cs^{n-2}\exp\left(-\frac{1}{2\sigma^2}[\sum (x_r-\bar{x})^2+n(\bar{x}-m)^2]\right) ds d\bar{x}. \qquad (3)$$

This represents the probability that a sample will be drawn from the population, having a mean lying between $\bar{x}$ and $\bar{x}+d\bar{x}$, and a standard deviation between s and $s+ds$. It follows that, given the standard deviation s, the law of distribution of samples of the means is represented by the normal curve

$$z = z_0 \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-m)^2\right) \qquad (4)$$

distributed about the same position as (1), but with standard deviation $\sigma/\sqrt{n}$.

In the same way, if we regard $\bar{x}$ as fixed, the law of distribution of the standard deviation of samples is given by

$$y = y_0\, s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right). \qquad (5)$$

The constant y_0 may be found as follows:

Let
$$I_p = \int\limits_0^\infty s^p \exp\left(-\frac{ns^2}{2\sigma^2}\right) ds$$
$$= \frac{\sigma^2}{n}\int\limits_0^\infty s^{p-1}\frac{d}{ds}\exp\left(-\frac{ns^2}{2\sigma^2}\right) ds$$
$$= \frac{\sigma^2}{n}(p-1)I_{p-2},$$
upon integration by parts. Hence we obtain
$$I_{n-2} = \left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}(n-2)}(n-3)(n-5)\ldots 1\, I_0,$$
or
$$I_{n-2} = \left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}(n-2)}(n-3)(n-5)\ldots 2\, I_1,$$
according as n is even or odd.

Evidently we have $I_0 = \sqrt{\left(\frac{\pi}{2n}\right)}\sigma$, and $I_1 = \frac{\sigma^2}{n}$. Since the area under the curve (5) represents the total frequency N of the population, we obtain

$$\left.\begin{aligned} y &= \frac{N}{(n-3)(n-5)\ldots 3.1}\sqrt{\frac{2}{\pi}}\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2}(n-1)} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) \\ &\text{when } n \text{ is even, and} \\ y &= \frac{N}{(n-3)(n-5)\ldots 4.2}\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2}(n-1)} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) \end{aligned}\right\} \quad (6)$$

when n is odd.

In a similar manner we obtain the value of the constant z_0 in (4). Denoting by x the distance of the mean of the sample from the mean of the original population, we have
$$N = z_0\int\limits_{-\infty}^{\infty}\exp\left(-\frac{nx^2}{2\sigma^2}\right) dx,$$
i.e.
$$N = 2z_0\sqrt{\left(\frac{\pi}{2n}\right)}\sigma.$$
Thus (4) can be written as
$$z = \sqrt{\frac{n}{2\pi}}\,\frac{N}{\sigma}\exp\left(-\frac{nx^2}{2\sigma^2}\right). \quad (7)$$

Now introduce the variable ζ defined by

$$\zeta = x/s. \tag{8}$$

The probability of obtaining a value of s lying between s and $s+ds$ is, by (5),

$$P_s = \frac{\int\limits_s^{s+ds} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) ds}{\int\limits_0^{\infty} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) ds} = \frac{y_0}{N} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) ds.$$

The probability of obtaining a value of x lying between x and $x+dx$ is, by (7),

$$z = \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{nx^2}{2\sigma^2}\right) dx = \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{ns^2\zeta^2}{2\sigma^2}\right)s\, d\zeta.$$

This is also the probability of obtaining a value of ζ between ζ and $\zeta+d\zeta$ for a given value of s.

Hence, the probability of obtaining a value of ζ between ζ and $\zeta+d\zeta$, while x lies between x and $x+dx$ and s between s and $s+ds$, is

$$\delta P = \frac{y_0}{N} s^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right) ds \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{ns^2\zeta^2}{2\sigma^2}\right)s\, d\zeta$$

$$= \sqrt{\frac{n}{2\pi}}\frac{y_0}{N\sigma} s^{n-1}\exp\left(-\frac{ns^2}{2\sigma^2}(1+\zeta^2)\right) ds d\zeta.$$

It follows that the probability of obtaining a value of ζ between ζ and $\zeta+d\zeta$ for any value of s is

$$P = \sqrt{\frac{n}{2\pi}}\frac{y_0}{N\sigma}\int\limits_0^{\infty} s^{n-1}\exp\left(-\frac{ns^2}{2\sigma^2}(1+\zeta^2)\right) ds d\zeta.$$

Hence, by (6), we obtain the results

$$\left.\begin{aligned} P &= \frac{1}{2}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{5.3}{4.2}(1+\zeta^2)^{-\frac{1}{2}n}\, d\zeta \quad (n \text{ odd}) \\ P &= \frac{1}{\pi}\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\frac{4.2}{3.1}(1+\zeta^2)^{-\frac{1}{2}n}\, d\zeta \quad (n \text{ even}) \end{aligned}\right\} \tag{9}$$

We note the very remarkable fact that the formula (9) *does not involve the unknown constant* σ: hence its practical importance.

Now write $\qquad \zeta = \tan\theta.$

Then, from (9), the probability of obtaining a value of ζ lying between ζ_1 and $-\zeta_1$, say, is

$$P = \frac{1}{2}\,\frac{n-2}{n-3}\,\frac{n-4}{n-5}\cdots\frac{5}{4}\cdot\frac{3}{2}\int\limits_{\tan^{-1}(-\zeta_1)}^{\tan^{-1}(\zeta_1)} \cos^{n-2}\theta\,d\theta,$$

or

$$P = \frac{1}{\pi}\,\frac{n-2}{n-3}\,\frac{n-4}{n-5}\cdots\frac{4}{3}\cdot\frac{2}{1}\int\limits_{\tan^{-1}(-\zeta_1)}^{\tan^{-1}(\zeta_1)} \cos^{n-2}\theta\,d\theta,$$

according as n is odd or even.

For further information on this subject, reference should be made to 'Student', *Biometrika* (1908), and R. A. Fisher, *Biometrika* (1914–15).

Using the above results we can construct a twofold table from which the significance of a variation of ζ between $\pm\zeta_1$ $\left(\text{i.e. of } \frac{\bar{x}-m}{s}\right)$ can be determined for a given value of n. For use in practice, Fisher has found it convenient to replace ζ by $t = \zeta\sqrt{n} = \frac{\bar{x}-m}{s}\sqrt{n}$. The values of t, P, and n are given in Table IV of Fisher's *Statistical Methods for Research Workers*, where it is to be noted that the n there used is less by unity than that taken above, and that m is assumed to be zero.

Other Tests for Significance

In the investigation given on p. 169 we have measured the *significance* of a pair of extractions by comparing the probability of obtaining such a pair with that of obtaining the 'most likely' pair. However, this is by no means the only method of estimating significance: consider, for example, the following problem.†

Suppose that we have a population of black and white balls in an unknown proportion $p:1$, and that we draw from it two samples, each consisting of 6 balls, which together contain 8 black balls. Thus, if the first sample contains r black balls,

Black	*White*
r	$6-r$
$8-r$	$r-2$

† Irwin, *Metron*, **12** (1935), 73.

the second will contain $8-r$ black balls. The probability of obtaining such a pair of samples is

$$^{6}C_{r}p^{r}(1-p)^{6-r}\,{}^{6}C_{8-r}p^{8-r}(1-p)^{r-2},$$

where r may assume all integral values from 2 to 6.

In the accompanying scheme we give the values assumed by the function

$$P(r) = {}^{6}C_{r}{}^{6}C_{8-r}$$

as r varies from 2 to 6.

r	$P(r)$
2	15
3	120
4	225
5	120
6	15

It follows that the probability of obtaining a table for which $r = 2$ is

$$15\Big/\sum_{r=2}^{6}P(r) = 15/495 = 1/33.$$

The probability of obtaining an equally probable or less probable table is $30/495 = 2/33$.

The same method may be employed when the two samples extracted are not of equal size. Thus, suppose that the first sample contains a fixed number $a+b$ of balls, and that the second contains a fixed number $c+d$, while the two samples together always contain $a+c$ black balls. In the table shown,

Black	*White*
$a+b-r$	r
$c-b+r$	$b+d-r$

the number r evidently cannot exceed the lesser of $a+b$ and $c+d$. The probability of obtaining such a table is

$$^{a+b}C_{r}\,{}^{c+d}C_{b+d-r}\big/{}^{a+b+c+d}C_{a+c}.$$

Consider, for instance, the following data, which give the number of cases of measles prevented and not prevented by the use of serum in each of two different schools.

	Prevented	*Not prevented*
School I	26	2
School II	61	2
Totals	87	4

The possible tables which may be enumerated are represented by the scheme

$28-r$	r
$59+r$	$4-r$

in which r may assume the values 0, 1, 2, 3, 4. We now calculate the corresponding values of the function

$$P(r) = {}^{28}C_r\, {}^{63}C_{4-r}.$$

These are shown in the accompanying table,

r	$P(r)$
0	595,665
1	1,111,908
2	738,234
3	206,388
4	20,475

from which it follows that

$$\sum_{r=0}^{4} P(r) = 2{,}672{,}670.$$

Hence, the probability of obtaining a table as improbable as or less probable than the observed one (for which $r = 2$) is

$$\frac{P(0)+P(2)+P(3)+P(4)}{\sum\limits_{r=0}^{4} P(r)} = \frac{1{,}560{,}762}{2{,}672{,}670} = 0{\cdot}584.$$

As a further illustration of how the significance of samples drawn from a population can be reduced to a comparison of relative probabilities we examine the following problem.†

Two populations each possess a certain quality in unknown proportions p_1 and p_2. Samples of magnitude N are drawn from each and found to contain x_1 and x_2 respectively of the quality in question. We inquire what is the significance relative to the

† See Jeffreys, *Proc. Camb. Phil. Soc.* **31** (1935), 203.

possible values of p_1 and p_2 to be attached to the difference x_1-x_2 found from the two samples. It is clear that the question becomes important only when x_1-x_2 is small compared with N.

Accordingly we examine the respective probabilities that samples x_1 and x_2 will be drawn from the two populations on the assumptions

(1) that $p_1 \neq p_2$,

(2) that $p_1 = p_2$.

In the case (1) the probability of x_1 and x_2 successes in N trials each is, by Bernoulli's Theorem,

$$A = {}^{N}C_{x_1}p_1^{x_1}(1-p_1)^{N-x_1}\times {}^{N}C_{x_2}p_2^{x_2}(1-p_2)^{N-x_2}$$

$$= \frac{N!^2}{x_1!\,x_2!\,(N-x_1)!\,(N-x_2)!}p_1^{x_1}(1-p_1)^{N-x_1}p_2^{x_2}(1-p_2)^{N-x_2}.$$

Now, prior to the drawing of the sample, p_1 and p_2 may have any values in the range

$$0 < (p_1, p_2) \leqslant 1,$$

all, it will be assumed, with equal probability.

Hence, on this basis the probability of drawing two samples x_1 and x_2 is

$$\int_0^1\int_0^1 A\,dp_1\,dp_2.$$

Now $$\int_0^1 dp\,p^m(1-p)^n = \frac{m!\,n!}{(m+n+1)!}.$$

Hence

$$\int_0^1 p_1^{x_1}(1-p_1)^{N-x_1}\,dp_1\int_0^1 p_2^{x_2}(1-p_2)^{N-x_2}\,dp_2$$

$$= \frac{x_1!\,x_2!\,(N-x_1)!\,(N-x_2)!}{(N+1)!\,(N+1)!}.$$

In the case (2), where $p_1 = p_2$, the probability of drawing samples x_1 and x_2 is

$$B = {}^{N}C_{x_1}p_1^{x_1}(1-p_1)^{N-x_1}\times {}^{N}C_{x_2}p_1^{x_2}(1-p_1)^{N-x_2}$$

$$= \frac{(N!)^2}{x_1!\,x_2!\,(N-x_1)!\,(N-x_2)!}p_1^{x_1+x_2}(1-p_1)^{2N-x_1-x_2}.$$

Again p may be assumed to range with equal probability between 0 and 1.

Hence the probability of drawing the two samples in this case is

$$\int_0^1 B\,dp_1.$$

Now

$$\int_0^1 p_1^{x_1+x_2}(1-p_1)^{2N-x_1-x_2} = \frac{(x_1+x_2)!\,(2N-x_1-x_2)!}{(2N+1)!}.$$

In A and B the coefficients not involving p_1 and p_2 are identical.

Hence we obtain the result:

$$\frac{\text{Probability of } (x_1, x_2) \text{ arising when } p_1 = p_2}{\text{Probability of } (x_1, x_2) \text{ arising when } p_1 \neq p_2}$$

$$= \frac{(x_1+x_2)!\,(2N-x_1-x_2)!}{(2N+1)!}\,\frac{\{(N+1)!\}^2}{x_1!\,x_2!\,(N-x_1)!\,(N-x_2)!}.$$

Assuming that N, x_1, and x_2 are all large numbers, by using Stirling's theorem we can approximate to this ratio; it becomes

$$\frac{N^{\frac{3}{2}}}{\sqrt{\{\pi(x_1+x_2)(2N-x_1-x_2)\}}}\exp\left\{-\frac{N(x_1-x_2)^2}{(x_1+x_2)(2N-x_1-x_2)}\right\}.$$

The problem of discriminating between the values of p for the two populations arises only when $\frac{x_1}{N}-\frac{x_2}{N}$ is small. For, by the method of Maximum Likelihood, x_1/N and x_2/N are the values of p_1 and p_2 for which x_1 and x_2 are the most probable samples. Accordingly let us write

$$\frac{1}{2}\left(\frac{x_1}{N}+\frac{x_2}{N}\right) = p$$

and

$$\frac{x_1}{N}-\frac{x_2}{N} = \delta p.$$

Then, finally, we may say that the relative likelihood equals

$$\frac{\text{Probability of drawing } (x_1, x_2) \text{ when the populations are identical}}{\text{Probability of drawing } (x_1, x_2) \text{ when the populations are different}}$$

$$= \left[\frac{N}{4\pi p(1-p)}\right]^{\frac{1}{2}}\exp\left\{-\frac{N\,\delta p^2}{4p(1-p)}\right\} = \frac{L}{\sqrt{\pi}}\exp(-L^2\,\delta p^2),$$

where

$$L^2 = \frac{N}{4p(1-p)}.$$

If d, the actual difference between the two successful drawings

x_1 and x_2, is constant for a population of increasing size, then clearly $L^2\,\delta p^2$ does not depend on N and the exponential term remains constant.

Thus, for a given difference $x_2 - x_1$ between the two readings, the relative likelihood that the two populations do not differ increases directly with L, i.e. with $N^{\frac{1}{2}}$. When $p = \frac{1}{2}$, $L = N^{\frac{1}{2}}$.

Ex. 1. For a given difference $x_1 - x_2$ in the samples of given size N, find the probability p for which the relative likelihood, that the populations are identical, is a maximum.

Ex. 2. Suppose that two samples, 100 each in number, are drawn from two bags containing black and white balls with the following results:

	Total	*Black*	*White*
First sample	100	41	59
Second sample	100	49	51

We have

$$N = 100, \quad p = \frac{1}{2}\left(\frac{41+49}{100}\right) = 0{\cdot}45,$$

$$\delta p = \frac{49-41}{100} = 0{\cdot}08,$$

$$L^2 = \frac{100}{4 \times 0{\cdot}45 \times 0{\cdot}55} = 100, \text{ approx.}$$

Thus

$$L = 10.$$

Hence

$$\frac{\text{Probability of identity}}{\text{Probability of difference}} = \frac{10}{\sqrt{\pi}}\exp(-100 \times 0{\cdot}08^2) = \frac{10}{\sqrt{\pi}}\exp(-0{\cdot}64),$$

i.e. it is approximately three times more probable that the two populations are identical than that they are different.

EXAMPLES ON CHAPTER IX

Ex. 1. The probability of landing a shot within the annulus of radii r and $r+dr$ on a target is $\frac{2h}{\sqrt{\pi}}\exp(-h^2r^2)\,dr$. A thousand shots are fired at the target and 500 are found to lie within 1 ft. of the centre. What is the number of shots expected to lie within 3 in. of the centre, and what is the least distance from the centre within which one shot is likely to be found?

Taking the unit of length as 1 ft., we have, by hypothesis,

$$\frac{2h}{\sqrt{\pi}}\int\limits_0^1 e^{-h^2r^2}\,dr = \tfrac{1}{2},$$

or $$\operatorname{Erf} h = 0{\cdot}5.$$

Thus $$h = 0{\cdot}48.$$

Hence the number of shots likely to be found within 3 in. of the centre is

$$\frac{1{,}000h}{\sqrt{\pi}}\int\limits_0^{\frac{1}{4}} e^{-h^2r^2}\,dr.$$

The least distance r from the centre within which one shot is likely to be found is given by

$$\frac{1{,}000h}{\sqrt{\pi}}\int\limits_0^{r} e^{-h^2r^2}\,dr = 1.$$

Ex. 2. *The Method of Least Squares.* On p. 169 we have derived this method from the normal law; the same result may, however, be obtained, without such an assumption, by introducing the concept of *weight*. Suppose that $x, y, z\ldots$ are n numbers and that $L_1 = a_1x+b_1y+c_1z+\ldots$, $L_2 = a_2x+b_2y+c_2z+\ldots,\ldots, L_s = a_sx+b_sy+c_sz+\ldots$ are s ($> n$) linear functions of $x, y\ldots$ with given coefficients, for which we have the estimated values $u_1, u_2,\ldots, u_s$. Then for the expression

$$L = \lambda_1 L_1+\lambda_2 L_2+\ldots+\lambda_s L_s,$$

where the λ's are constant, we shall have the estimate

$$\lambda_1 u_1+\lambda_2 u_2+\ldots+\lambda_s u_s.$$

If we choose the λ's so that

$$\lambda_1 a_1+\lambda_2 a_2+\ldots+\lambda_s a_s = 0, \qquad \lambda_1 b_1+\lambda_2 b_2+\ldots+\lambda_s b_s = 0, \quad \text{etc.},$$

then L will reduce to x, in which case $\lambda_1 u_1+\lambda_2 u_2+\ldots+\lambda_s u_s$ is an estimated value for x.

We now define the weight W of L, for any set of values of $\lambda_1, \lambda_2,\ldots$, by the expression

$$\frac{1}{W} = \lambda_1^2+\lambda_2^2+\ldots+\lambda_s^2.$$

Further, we assume that the best estimate for x is that for which W is a maximum. This gives us the condition $\lambda_1\,d\lambda_1+\lambda_2\,d\lambda_2+\ldots+\lambda_s\,d\lambda_s = 0$.

If we solve the equation so obtained, using the method of undetermined multipliers, we find for x the value that would have been derived from the Method of Least Squares as formerly explained.

For further details, as well as for justification of the present assumptions, the reader may consult Whittaker and Robinson, *The Calculus of Observations*, § 115.

Ex. 3. The speed of a train is recorded every second by an instrument which in reality gives the average reading over the previous T seconds. If the recorded speed $u(t)$ is found to follow the formula

$$u(t) = at^2+bt+c,$$

determine the true speed $v(t)$.

Here

$$
\begin{aligned}
u(t) &= \frac{1}{T}\int\limits_{-T}^{0} v(t+x)\,dx = \frac{1}{T}\int\limits_{-T}^{0} e^{xD}v(t)\,dx, \quad \text{where} \quad D = \frac{d}{dt}, \\
&= \frac{1}{T}\left[\frac{e^{xD}}{D}\right]_{-T}^{0} v(t) = \frac{1}{TD}[1-e^{-TD}]v(t) \\
&= \frac{1}{TD}\left(TD-\frac{T^2D^2}{2!}+\frac{T^3D^3}{3}-\ldots\right)v(t) \\
&= \left(1-\frac{TD}{2}+\frac{T^2D^2}{6}-\ldots\right)v(t).
\end{aligned}
$$

Hence

$$
\begin{aligned}
v(t) &= \left(1-\frac{TD}{2}+\frac{T^2D^2}{6}-\ldots\right)^{-1}u(t) \\
&= \left(1+\frac{TD}{2}+\frac{T^2D^2}{12}+\ldots\right)u(t) \\
&= u(t)+\frac{T}{2}u'(t)+\frac{T^2}{12}u''(t)\ldots \\
&= u\left(t+\frac{T}{2}\right)-\frac{T^2}{24}u''(t), \quad \text{approximately.}
\end{aligned}
$$

Hence

$$
\begin{aligned}
v(t) &= at^2+bt+c+\frac{T}{2}(2at+b)+\frac{T^2}{12}2a \\
&= at^2+(b+aT)t+\frac{aT^2}{6}+\frac{bT}{2}+c.
\end{aligned}
$$

Ex. 4. Show that, if $u(t)$ is the recorded measurement at time t, where in fact it is the average over a period $2T$ lying evenly about time t, then the true measurement $v(t)$ is given by

$$
\begin{aligned}
v(t) &= u(t)-\tfrac{1}{6}T^2u''(t)+\tfrac{7}{360}T^4u^{\text{iv}}(t)\ldots \\
&\doteq u(t)-\tfrac{1}{6}T^2\Delta^2u(t)+\tfrac{1}{6}T^2\Delta^3u(t)+\tfrac{7}{360}\Delta^4u(t)\ldots.
\end{aligned}
$$

0	0	0	0	0	0
0	1·5	1·8	1·8	1·6	0
0	1·8	2·1	1·9	1·6	0
0	1·8	2·2	2·0	1·7	0
0	1·6	2·0	2·2	1·7	0
0	0	0	0	0	0

Ex. 5. In a square lake depth soundings are taken from a boat at a series of points forming the corners of the 25 squares into which the surface of the lake is divided. The errors in placing the boat in each position for sounding are given by the law Ae^{-r^2}, where r is the accidental deviation from the true position. If the figures in the diagram are the readings obtained, find the true distribution of depth.

Ex. 6. Show that, if different sets of observations each satisfy a Gaussian law of the form $\frac{h}{\sqrt{\pi}}e^{-h^2x^2}$, where the h's may adopt any values occurring with a probability given by $\frac{\lambda}{\sqrt{\pi}}e^{-h^2a^2}$, then the law of distribution of the x's is of the form $\frac{\lambda}{\pi(x^2+\lambda^2)}$.

Ex. 7. For a probability distribution of the type $\frac{2\lambda^3}{\pi}\frac{1}{(x^2+\lambda^2)^2}$, find the mean value of x and x^2.

APPENDIX

$$\mathrm{Erf}\, x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

x	Erf x	x	Erf x	x	Erf x
0·02	0·02256	1·02	0·85084	2·02	0·99572
0·04	0·04511	1·04	0·85865	2·04	0·99609
0·06	0·06762	1·06	0·86614	2·06	0·99642
0·08	0·09008	1·08	0·87333	2·08	0·99673
0·10	0·11246	1·10	0·88021	2·10	0·99702
0·12	0·13476	1·12	0·88679	2·12	0·99728
0·14	0·15695	1·14	0·89308	2·14	0·99753
0·16	0·17901	1·16	0·89910	2·16	0·99775
0·18	0·20093	1·18	0·90484	2·18	0·99795
0·20	0·22270	1·20	0·91031	2·20	0·99814
0·22	0·24430	1·22	0·91553	2·22	0·99831
0·24	0·26570	1·24	0·92051	2·24	0·99846
0·26	0·28690	1·26	0·92524	2·26	0·99861
0·28	0·30788	1·28	0·92973	2·28	0·99874
0·30	0·32863	1·30	0·93401	2·30	0·99886
0·32	0·34913	1·32	0·93807	2·32	0·99897
0·34	0·36936	1·34	0·94191	2·34	0·99906
0·36	0·38933	1·36	0·94556	2·36	0·99915
0·38	0·40901	1·38	0·94902	2·38	0·99924
0·40	0·42839	1·40	0·95229	2·40	0·99931
0·42	0·44747	1·42	0·95538	2·42	0·99938
0·44	0·46623	1·44	0·95830	2·44	0·99944
0·46	0·48466	1·46	0·96105	2·46	0·99950
0·48	0·50275	1·48	0·96365	2·48	0·99955
0·50	0·52050	1·50	0·96611	2·50	0·99959
0·52	0·53790	1·52	0·96841	2·52	0·99963
0·54	0·55494	1·54	0·97059	2·54	0·99967
0·56	0·57162	1·56	0·97263	2·56	0·99971
0·58	0·58792	1·58	0·97455	2·58	0·99974
0·60	0·60386	1·60	0·97635	2·60	0·99976
0·62	0·61941	1·62	0·97804	2·62	0·99979
0·64	0·63459	1·64	0·97962	2·64	0·99981
0·66	0·64938	1·66	0·98110	2·66	0·99983
0·68	0·66378	1·68	0·98249	2·68	0·99985
0·70	0·67780	1·70	0·98379	2·70	0·99987
0·72	0·69143	1·72	0·98500	2·72	0·99988
0·74	0·70468	1·74	0·98613	2·74	0·99989
0·76	0·71754	1·76	0·98719	2·76	0·99991
0·78	0·73001	1·78	0·98817	2·78	0·99992
0·80	0·74210	1·80	0·98909	2·80	0·99992
0·82	0·75381	1·82	0·98994	2·82	0·99993
0·84	0·76514	1·84	0·99074	2·84	0·99994
0·86	0·77610	1·86	0·99147	2·86	0·99995
0·88	0·78669	1·88	0·99216	2·88	0·99995
0·90	0·79691	1·90	0·99279	2·90	0·99996
0·92	0·80677	1·92	0·99338	2·92	0·99996
0·94	0·81627	1·94	0·99392	2·94	0·99997
0·96	0·82542	1·96	0·99443	2·96	0·99997
0·98	0·83423	1·98	0·99489	2·98	0·99997
1·00	0·84270	2·00	0·99532	3·00	0·99998
				3·123	0·999990
				3·459	0·999999

INDEX

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, OXFORD
BY CHARLES BATEY, PRINTER TO THE UNIVERSITY